DESIGN METHODOLOGY FOR RELIABLE AND ENERGY EFFICIENT SELF-TUNED ON-CHIP VOLTAGE REGULATORS

A Dissertation
Presented to
The Academic Faculty

By

Venkata Chaitanya Krishna Chekuri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2021

DESIGN METHODOLOGY FOR RELIABLE AND ENERGY EFFICIENT SELF-TUNED ON-CHIP VOLTAGE REGULATORS

Approved by:

Dr. Saibal Mukhopadhyay, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Abhijit Chatterjee
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Sung Kyu Lim
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Tushar Krishna
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Hyesoon Kim
School of Computer Science
Georgia Institute of Technology

Date Approved: January 21, 2021

Out of clutter, find Simplicity. From discord, find Harmony. In the middle of difficulty lies Opportunity.

Albert Einstein

To my family

ACKNOWLEDGEMENTS

My PhD years have been one of the most formative phases of my life and I would like to take this opportunity to express my gratitude for the people who have contributed to and encouraged me in this work. I would like to acknowledge and express the highest gratitude to my advisor, Prof. Saibal Mukhopadhyay. He took a chance on me when I needed it and has been a constant source of inspiration and support to try out new ideas and push the boundaries of research. I am thankful to him for his motivation and guidance on this project which I started without any background knowledge and providing the resources necessary to succeed. From him, I have gained not only technical expertise, but also an ability to guide and mentor others. I am also grateful to him for never letting me worry about the funding and providing an opportunity to travel all around the USA and the world for conferences. I would like to thank Prof. Abhijit Chatterjee, Prof. Sung-Kyu Lim, Prof. Tushar Krishna and Prof. Hyesoon Kim for being part of my thesis committee. Their valuable inputs played a major role in shaping my final thesis.

I am grateful to Monodeep Kar, Arvind Singh, Nael Rahman and Edward Lee who were my collaborators for multiple projects and demonstrating a sincere interest on success of the projects. I would like to thank SRC and DARPA for providing funding for projects that contributed towards my thesis.

I was fortunate enough to spend a semester during my PhD at Nvidia. I would like to thank my mentors, Santosh and Semmal Ganapathy for allowing me to work on cutting edge research, providing important technical and professional knowledge and teaching me how to present ideas in a simple yet effective manner. I would also like to take this opportunity to thank Prof. Sung-Kyu Lim and Prof. Tushar Krishna for collaborating on a research project and teaching me topics outside my area of expertise.

I would like to thank all members in GREEN Lab, Georgia Tech for their guidance, friendship, and support. Monodeep Kar, Arvind Singh and Faisal Amir for introducing

me to the lab, starting me off on the right foot and the many wonderful coffee breaks and meals. I would like to thank all the present and past Green Lab members for their valuable discussions and insights; Jaeha Kung, Duckhwan Kim, Faisal Amir, Jong Hwan Ko, Taesik Na, Yun Long, Burhan Mudassar, Nael Mizanur Rahman, Edward Lee, Nihar Dasari, Daehyun Kim and all other members. Especially, I am thankful to Monodeep Kar and Arvind Singh for being the best mentors I could ask for and Nael Rahman and Edward Lee for being a great teammates for many research projects. I will never forget the moments spent with my colleagues and I look forward to meeting them in future again.

I want to thank all of my teachers at Georgia Tech, whose classes helped me better understand the various topics in Electrical and Computer Engineering. I would also like to take this opportunity to thank Keith May, Pamela Halverson, Faith Midkiff, Daniela Staiculescu, and Tasha Torrence for their technical and administrative support.

I would like to thank all my close friends in Atlanta and the US, colleagues in the Klaus building, CRC buddies, friends from my undergrad days and the people I met during my internships, for all the amazing and memorable times. Thanks for the wonderful meals, trips, gatherings, and stimulating discussions. I have had a special rapport with each one of you personally and I cherish that bond.

Last but foremost, I would like to thank my parents and my brother for the unconditional love and support they have given me over the course of my life. I wouldn't have made it this far without their support during the course of the PhD, particularly at tough times during tapeouts and deadlines. It is the difficult decisions and sacrifices that they have made, that has brought me to the position I am in right now, and for that I am eternally thankful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The energy-efficiency needs in computing systems, ranging from high performance processors to low-power devices is steadily on the rise, resulting in increasing popularity of on-chip voltage regulators (VR). The high-frequency and high bandwidth on-chip voltage regulators such as Inductive voltage regulators (IVR) and Digital Low Dropout regulators (DLDO) significantly enhance the energy-efficiency of a SoC by reducing supply noise and enabling faster voltage transitions. However, IVRs and DLDOs need to cope with the higher variability that exists in the deep nanometer digital nodes since they are fabricated on the same die as the digital core affecting performance of both the VR and digital core. Moreover, in most modern SoCs where multiple power domains are preferred, each VR needs to be designed and optimized for a target load demand which significantly increases the design time and time to market for VR assisted SoCs.

This thesis investigates a performance-based auto-tuning algorithm utilizing performance of digital core to tune VRs against variations and improve performance of both VR and the core. We further propose a fully synthesizable VR architecture and an auto-generation tool flow that can be used to design and optimize a VR for given target specifications and auto-generate a GDS layout. This would reduce the design time drastically. And finally, a flexible precision IVR architecture is also explored to further improve transient performance and tolerance to process variations. The proposed IVR and DLDO designs with an AES core and auto-tuning circuits are prototyped in two testchips in 130nm CMOS process and one test chip in 65nm CMOS process. The measurements demonstrate improved performance of IVR and AES core due to performance-based auto-tuning. Moreover, the synthesizable architectures of IVR and DLDO implemented using auto-generation tool flow showed competitive performance with state of art full custom designs with orders of magnitude reduction in design time. Additional improvement in transient performance of IVR is also observed due to the flexible precision feedback loop design.

CHAPTER 1
INTRODUCTION

With ever increasing number of integrated circuits, voltage regulators have become a critical component of any design. Power consumption has become one of the most important issues in modern silicon on chips (SoCs). Absence of these voltage regulators can prove to be fatal in most high frequency and high performance circuit designs. As a result, on-chip integrated voltage regulators (IVRs) including fully integrated inductive VRs (FIVR) with on-chip/on-package passives and low-dropout (LDO) VRs are becoming an integral part of modern digital processors. The high-frequency/high bandwidth on-chip voltage regulators significantly enhance the energy-efficiency of a SoC by reducing supply noise and enabling faster voltage transitions.

However, with the benefits come the challenges of achieving those features with minimal possible changes to the designs themselves. IVRs need to cope with the higher variability that exists in the deep nanometer digital nodes. An IVR's characteristics can shift due to process-induced variations in transistors' and passives' characteristics. The integrated (on-chip/on-package) passives (L/C) and their Equivalent Series Resistance (ESR) can have higher variability than off-chip components. The transistor variations impact the integrated power FET as well as characteristics of the control circuits, for example, delay variations in the compensator. The high ($>$100MHz) operating frequency of IVRs can make them susceptible to new sources of variations such as static/slow frequency drifts and jitter in the pulse train. Due to close proximity of IVR and core, the IVR's temperature can couple with core power due to thermal coupling. The aging of the power FETs and passives need to be considered as well which can also be accelerated due to higher temperature. Therefore, there is a need for on-line testing and self-tuning of high frequency IVR's to enable reliable operation of a digital processor.

Due to increasing power domains in modern SoCs, the VRs for each of the domains need to be individually designed and optimized for a target load resulting in increase in design time and complexity for the VR assisted SoCs. Thus, a generic architecture and associated methodology for automated design and physical layout generation of on-chip VRs, and integration of the generated VR within a digital SoC is needed to overcome the design time bottleneck.

In this thesis our primary goal is to develop energy efficient and robust on-chip voltage regulators, mainly inductive buck regulators and digital low-dropout (DLDO) regulators integrated in the same chip/package with a System-on-Chip (SoC). More specifically, we will focus on developing self tuning circuits for VRs to improve transient performance and an automated tool flow for fast VR design.

1.1 Problem Statement

The objective of the proposed research is to develop a robust design methodology for reliable and energy efficient self tuned on-chip voltage regulators, namely inductive integrated voltage regulators (IVR) and digital low dropout regulators (DLDO). This includes:

- Developing architectures and algorithms for a lightweight self tuning engines for improved transient performance against process and passive variations
- Exploring reliability aspects of the different on-chip voltage regulators to study the effects of voltage stress on transient performance and efficiency.
- Developing a specification to GDSII layout automated tool flow for on-chip voltage regulators to reduce the overall design time while optimizing for target load.
- Designing a fully synthesizable and flexible precision VR architecture to facilitate easy integration with the auto-generation tool flow and improve transient performance.

1.2 Key Contributions

The key contributions and findings of this thesis can be summarized as:

- **Improving performance of digital core using on-chip auto-tuning:** A performance based auto-tuning algorithm to tune a system of an IVR driving a digital core is implemented. In the proposed approach, using the performance of the digital core allows us to capture effect of process variations on chip along with the variations in passives. Thus, using performance based tuning we can enhance the digital system performance which is beyond the capability of the existing IVR tuning methods ([1, 2, 3]) as they do not consider performance of digital core in the tuning metric.

- **Analyzing effects of aging related degradations in on-chip voltage regulators:** The effects of NBTI induced aging degradations in on-chip VRs, namely IVR and DLDO have been analyzed. The effect of aging is explored in two locations: the power stage and digital control loop. For power stage aging, based on the dependency of closed loop transfer function on the PFET on-resistance, it is observed that DLDO is more susceptible to significant degradation in the transient performance whereas IVR has marginal effect on transient performance. However, IVR does undergo small drop in efficiency. Moreover, for the DLDO the degraded transient response can be improved using on-chip auto-tuning by adjusting the compensator gains. This analysis for on-chip VRs has been validated in silicon for the first time to best of our knowledge. Additionally, for the digital controller aging, it is observed that both VRs have significant degradation in transient response as the controller becomes slower and requires reducing the sampling frequency to operate.

- **Reducing the design time of an on-chip voltage regulator by orders of magnitude using an automated tool flow:** A scalable EDA tool flow for fast GDSII generation of on-chip VRs (IVR and DLDO) has been developed. Unlike the prior works [4, 5] in this area which are mainly used for low frequency analog controller based

off-chip VR, the proposed tool generates digitally controlled high-bandwidth VRs which have been validated in silicon. The proposed tool combines use of front end efficiency and time/frequency domain Simulink models along with a back end physical design flow. The front and back end flows are guided using an optimization flow that optimizes the control loop and power stage of the VR to achieve desired transient performance and/or efficiency for the target specifications. The auto-generated VR shows comparable performance with full/semi custom designs, while enabling orders of magnitude reductions in design time which would reduce time to market for VR-assisted SoCs.

- **Tolerating variations in control loop and improving transient performance using flexible precision VR architecture:** A fully synthesizable flexible precision and variable frequency feedback loop architecture is developed to improve the versatility of on-chip VR by enabling trading off accuracy (output quality) with transient response time. This feature enables tolerating variations in the VR control loop which typically requires reduction in sampling/switching frequencies of VR to ensure timing closure of the degraded control loop and resulting in slower transient response. However, the timing closure for the degraded control loop can be achieved by reducing the precision of feedback loop macros without reducing frequency resulting in better transient response. Moreover, it is observed that dynamically changing the precision and frequency of the control loop can be used as a form of non-linear control to improve the transient performance by sampling at faster rate and lower precision during transient events and slower rate and higher precision at steady state. And unlike prior works [1, 6] the fully synthesizbale and flexible precision feedback loop macros make it easy for the design to scale across process nodes and integrate with auto-generation tool flows.

1.3 Organization of this thesis

Chapter 2 provides a detailed literature survey on several topics which are essential to comprehend the scope and contributions of this thesis. This includes on-chip voltage regulators and effect of variations on the regulators. Existing technologies and techniques for tuning the voltage regulator against variations also discussed along with and mixed signal design automation tool flow for faster design time.

Chapter 3 discusses an auto-tuning method for IVR driven by the performance of the digital cores. A detailed simulation framework is developed and key simulation results are analyzed and discussed in this chapter.

Chapter 4 demonstrates the proposed performance based tuning from previous chapter in a 130nm CMOS test-chip. Overall system architecture including the hardware translation of the auto-tuning engine and key measurement results are also discussed in this chapter.

Chapter 5 explores the effects of negative bias temperature instability (NBTI) induced ageing on on-chip voltage regulators. Modelling of IVR and DLDO along with simulation and measurement setups in 130nm and 65nm process are focused upon in this chapter.

Chapter 6 provides a discussion on a specification-to-GDS layout auto-generation tool for on-chip voltage regulators. This includes a detailed look into tool flow consisting of front end transient and efficiency models and back end physical design flows guided by an optimization function. Capabilities of the proposed tool are discussed using case studies.

Chapter 7 discusses an all-digital, synthesizable, flexible precision and modular IVR architecture along with a synthesizable DLDO architecture. This includes a detailed look at the modular and synthesizable feedback loop macro architecture along with a macro generation tool flow. The flexible precision operation, IVR and DLDO are demonstrated in 65nm test-chip and key measurement results are discussed.

Chapter 8 highlights the key contributions of this dissertation and discusses future directions for the research.

CHAPTER 2

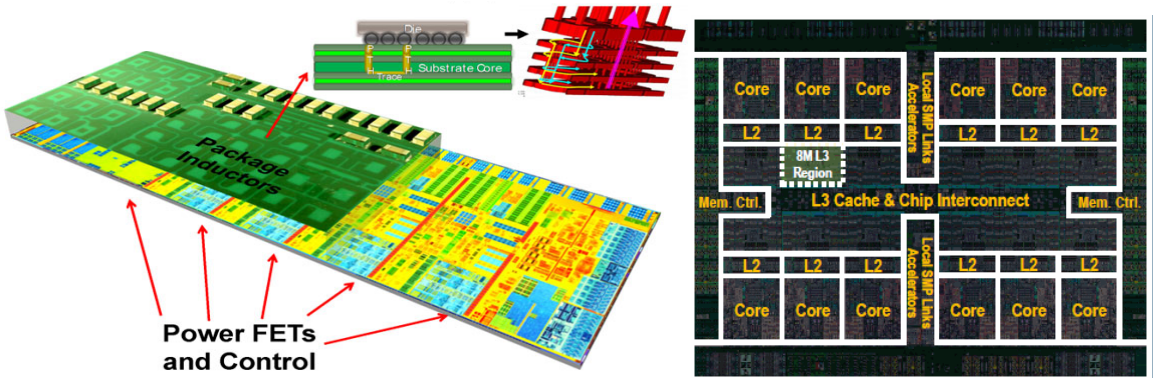
LITERATURE SURVEY

2.1 On-chip Power Management

2.1.1 Fully Integrated Voltage Regulator (FIVR)

Integration of inductive voltage regulators with on-chip/on-package inductors/capacitors on the same chip as the digital logic cores has received significant attention in recent years for designing power-efficient SoCs [9, 1, 10, 11, 7, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. The fully integrated inductive voltage regulators (Fig. 2.1a) for digital systems have demonstrated high loop bandwidth (>50MHz) to ensure a fast recovery from voltage droops due to load transition (load transient) and a fast transition of the output voltage (reference transient) to support dynamic voltage frequency scaling (DVFS).

IVRs improve the energy efficiency of a digital system by allowing fast recovery from transient droops as well as fast voltage ramp-rate during power-state change [1, 7, 24, 25]. Use of fully integrated inductive IVR have been on the rise for commercial high performance processors due to efficient integration of package/on-chip inductance as demonstrated in [7, 16]. A generic inductive IVR uses a power stage composed of a PMOS and NMOS device (on multiple of them depending on the voltage rating of the devices) and an on-die/on-package inductor and an on-chip output capacitor. A high switching frequency (>100MHz) is required to manage ripple with small L/C. Multiple phases of IVR can be used to reduce the voltage ripple, however increases the number of inductors in the system. A voltage mode PWM control is typically used as controller for these FIVRs. Due to the ease of integration into the advanced process nodes as well as high bandwidth due high operating clock frequency, digital PID compensators are preferred [1, 14, 16]. A type-III compensator with two zeros is required for compensating the filter double pole as the zero



(a) (b)

Figure 2.1: Modern processors implementing on-chip voltage regulation. (a) 4th generation Intel® Core™ Microprocessor [7] (b)IBM Power8™[8]

created by the ESR of the output capacitance resides at a high frequency. To improve the loop bandwidth which dictates the response speed from transient, phase shifting of sampling clocks [14] as well as reduced precision multi-sampling [1] have been used.

2.1.2 Digital Low Drop Out Regulators (DLDO)

IVRs consume significant on-chip resources mainly due to passives. To eliminate the need for large passives, digital low drop out (DLDO) regulators are preferred. DLDOs are well-known for easy implementation and fast transient response. Primary source of power loss in DLDOs are power stage losses. These are determined by the dropout voltage across the power stage and are prominent at lower output voltages. Thus, for large systems, DLDOs are used along with IVRs for more efficient and fine-grained point of load power-management (Fig. 2.1b and [8]). A DLDO can be implement in multiple ways. The most generic architecture includes being implemented using a power transistor array and controlled by a digital controller. One of the common control scheme includes utilizing shift-register (SR) based bang-bang control as presented by in Nasir et. al in [17, 26]. This architecture is compact but suffers from poor transient performance. For improved transient performance, recent digital LDO architectures either have an additional loop (analog-assisted) [27, 28] or proportional-integral-derivative (PID) controller [29, 18].

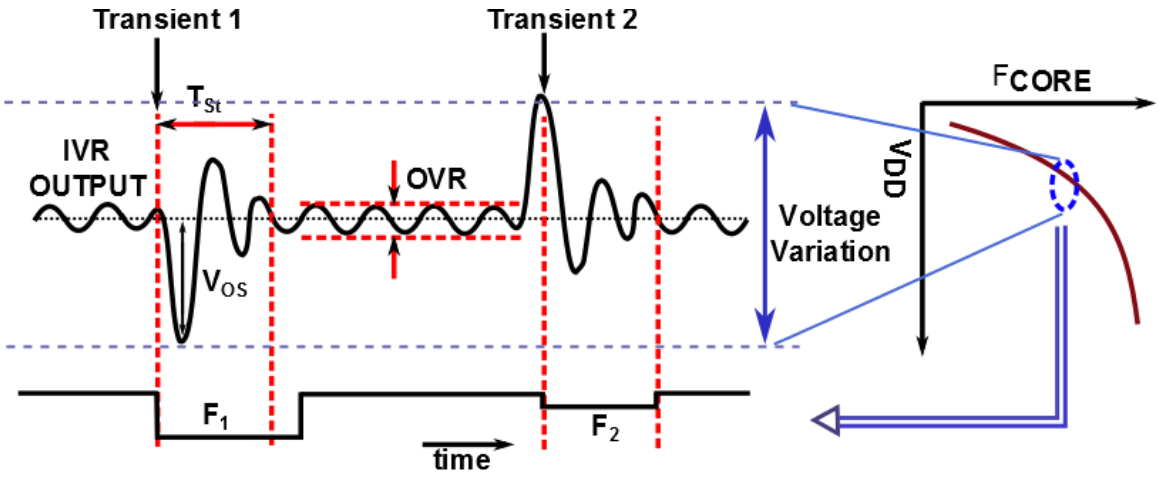


Figure 2.2: Performance dependency of digital cores on power supply quality

2.2 Circuit performance under variations

A critical challenge in designing circuits in nanometer digital process node is to tolerate process variation that affects performance of the digital circuits [30, 31, 32, 33]. As the IVRs are designed in the same process nodes, they are also expected to suffer from variations [1, 15, 34, 35]. In particular, on-chip/on-package passives are expected to suffer from higher variation than the off-chip discrete components [1, 34], resulting in variations in transient (load and reference) performance. The variations in the IVR's output response translates to increased power supply noise, which is further coupled to transistor variation, resulting in higher uncertainty in the performance of the digital cores (Fig. 2.2).

The performance of digital cores is determined by the shifts in the process (threshold voltage) as well as variation in the supply voltage. The supply variation is defined by the steady state perturbations as well as transient supply droop due to sudden current demand by a digital block. For IVRs the steady state perturbations are contributed by the output voltage ripple whereas the droop is dictated by the transient response of the control loop. Any variations in the passive values will change transient response and hence, supply noise experienced by the digital core. Moreover, the sensitivity of delay to supply voltage depends on the threshold voltage (higher sensitivity at higher V_t). Consequently, tuning of

IVR's coefficients directly based on the delay can account for the coupled effects of process variation and transient supply noise. Improving tolerance to supply/process variations helps reduce the voltage margin normally added in digital cores, and hence, improve maximum operating frequency and/or reduce power dissipation.

2.3 Supply dependency of digital circuits

Voltage scaling and/or frequency scaling techniques have been developed as possible effective solutions to improve performance and energy efficiency in the presence of dynamic variations. In high volume manufacturing (HVM), binning the processors under process variation involves a post-silicon tuning step where the minimum VCC is found out for a digital core under process variation to meet target frequency [30]. This process addresses the die-to-die process variations. To counter for within-die variations, spatial voltage-delay profile across a chip is tracked using replica circuits which are also useful to capture the effect of local supply droops as well as temperature fluctuations. With the IVRs distributed across the die for state-of-the-art processors, local digital blocks can be controlled with their individual supply voltages.

Apart from tuning against these static variations, digital logics in a high performance processor experience run-time dynamic variations like supply droop, coupling noise and temperature fluctuations. Two distinct approaches exists to tackle these effect:

- A voltage margin is added to the intrinsic minimum operating voltage of a digital circuit to meet the target frequency. The voltage margins are set pessimistically reducing the energy-efficiency of the system, but ensures error-free operation.

- Error tolerant designs like Razor [36, 37, 38] use aggressively smaller supply margin with a higher target frequency and use special circuits like error-detection flip-flops/latches to recover from runtime timing errors and improves energy efficiency.

The performance (defined as the number of instructions executed over a given time pe-

riod with functional correctness) of an error tolerant system is defined by the perturbations in the power supply. The supply quality is defined by the steady state perturbations as well as transient supply droop which is induced due to sudden current demand by a digital block sharing the same supply rail. For digital cores which are supplied by an off-chip voltage regulator module (VRM), the supply quality is determined by the local decupling capacitance as well as impedance of the power distribution network (PDN). For IVRs the steady state perturbations are contributed by the output voltage ripple whereas the transient droop is dictated mostly by the control loop. Any variations the passive values will change the IVR output quality and affect system performance. The tuning knobs to control these effects are the switching frequency and the loop compensation. Increasing switching frequency leads to reduction in power efficiency and therefore is not usually used as a control knob, leaving the loop compensation as the key control knob to tune the IVR. The sensitivity of the IVRs to variation in passive is higher than the off-chip VRMs [34], showing the need for auto-tuning in state-of-the-art microprocessors.

2.4 Prior Work on Auto-tuning Algorithms

Auto-tuning process for any VRMs (including IVRs) observes the output behavior of the VRM after perturbing the control loop and adjusts the controller transfer function based on a cost (Fig. 2.3). The post-silicon tuning of low frequency ($<$1MHz) off-chip VRs have been explored in the past. The existing techniques aim to directly characterize the IVR's frequency response (such as unity gain frequency, phase margin etc.) and steady state parameters. However, the tuning schemes are complex, require significant computation and memory, and difficult to scale to high frequency ($>$100MHz) IVRs. For example, the auto tuning algorithm presented by Shirazi et al [3] required 28,000 logic gates, four 1024 x 18-bit RAM blocks to compute and store the frequency response, one 256 x 16-bit ROM block for the complex exponential lookup table (LUT), and one 512 x 16-bit ROM block for the discrete-zero LUT (implemented in FPGA). In [39, 40], Costabber et al demonstrate

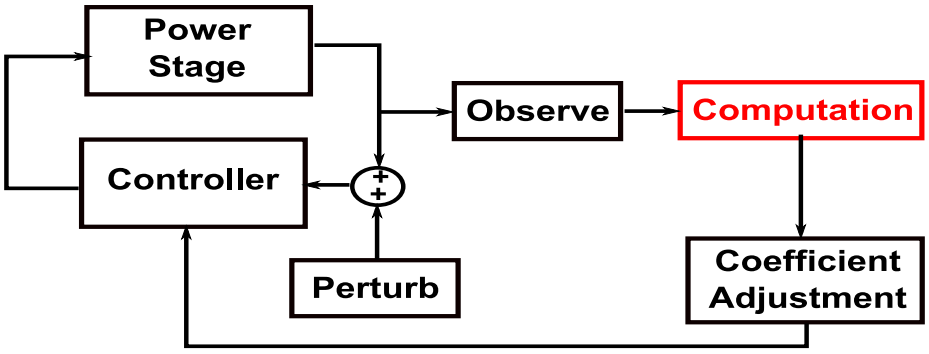


Figure 2.3: Traditional auto-tuning approach for swicthed regulators

an auto-tuning controller scheme based on model reference impulse response. This auto-tuning controller compares the measured system response with a reference system response and adjusts a compensator parameter accordingly to minimize the error function. In [41], Stefanutti et al present an autotuning controller based on the relay feedback method. It tunes the proportional–integral–derivative (PID) parameters of the compensator based on a desired phase margin and control loop bandwidth. In [42], Saggini et al propose a self-tuning analog current-mode controller. The tuning is based on the insertion of nonlinear blocks in the control loop and measurement of the closed-loop properties such as gain margin, phase margin, and crossover frequency by perturbing the output voltage. The controller is then tuned according to the desired set of specifications.

Most of frequency-domain tuning algorithms discussed thus far involve FFT computation and uses complex computation engine. Although ideal for off-chip low frequency VRMs, these computation heavy algorithms become challenging to implement in high switching frequency IVRs. Time domain based tuning algorithm tunes the controller by performing simple arithmetic computations on the time domain samples of the IVR output instead of performing frequency domain analysis [1, 2]. In [1], Kar et al use a cost metric which is a summation of aggregated absolute error values, aggregated signed error values and settling time to a load transient which is induced in the middle of an evaluation cycle. In [2], Qahouq et al use the compensated error value to tune the coefficients, but it is only suitable for implementation in low frequency VRMs. In contrast to the prior work,

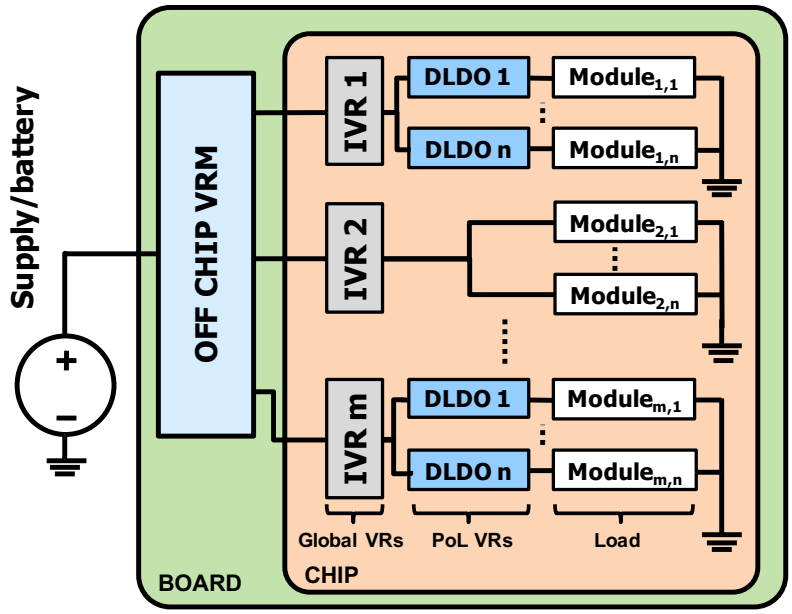


Figure 2.4: Example of distributed power domain in modern SoCs

this work proposes to tune the IVR's coefficients based on the delay of a digital circuit to simultaneously consider effects of passive and process variations.

2.5 Fully synthesizable voltage regulators

The moderns SoCs have many power domains (Fig. 2.4). The high-frequency/high-bandwidth on-chip voltage regulators for each power domain can significantly enhance the energy-efficiency of the chip. First, on-chip VRs reduce supply noise, thanks to fast response and less voltage droop due to load transition [9, 1, 10, 7]. Second, they enable faster voltage transitions enabling localized dynamic voltage frequency scaling (DVFS) [9, 1, 10, 7]. However, the controller and power stages of the VRs for each voltage domain must be independently designed to match the target load demand i.e. maximum steady state power, power quality (voltage ripple) and transient (load/reference) performance. Since voltage regulators are typically a mixed-signal design, they usually need manual optimization and custom layout thereby increasing the design time and delaying time to market for SoC requiring on-chip voltage regulators. An EDA tool for automated design and GDSII generation of an on-chip VR (such as DLDO and IVR), and integration of the generated

IVR/DLDO within a digital SoC will significantly reduce the design time of VR-assisted SoCs.

There have been some prior works [6, 43, 44] for analog/mixed signal design automation which use custom cells along with foundry provided standard cell to design circuits like PLLs and ADCs. But only a few prior works [4, 5] have been reported for on-chip regulators. Automated synthesis design flow of power converter blocks using equation and the simulation-based methods is discussed in [4] and [5], who demonstrate an automated design flow from specification to layout for DC-DC buck converter designs with current-mode controller. However, these works do not consider co-design of the controller models and the physical design of power stage and controller. Moreover, they implement a small analog controller at low switching frequencies ($<$10MHz), while high-frequency digital control is often preferred for modern SoCs. Finally, they do not discuss the integration of the VRs with digital core.

In recent years, the popularity of digitally controlled IVRs and DLDOs has increased. The digital nature of the controller makes it more appealing option as the design is less complicated and time-consuming as it can be designed in a Register-Transfer-Language (RTL) functional code and synthesized using standard place & route tools. For reference in [6], Choi et al presents a 4-phase IVR with digital adaptive on time control. For this design some parts of the control loop are synthesizable such as offset-controlled comparator and Adaptive ON time generator. However, there is insufficient information provided regarding the other parts traditionally analog and mixed signal blocks such are Digital to analog converter (DAC), VSW sensor, Delay lines and power stage and therefore are expected to be custom/manually designed and laid out. Thus, adding more synthesizable elements into the VR architecture makes it easier to design, scale and integrate into an auto-generation tool to reduce design time significantly.

CHAPTER 3

PERFORMANCE BASED AUTO-TUNING OF ALL DIGITAL FIVR

Integration of inductive voltage regulators with on-chip/on-package inductors on the same chip as the digital logic cores helps fast recovery from voltage droops (load transient) and a fast transition of the output voltage (reference transient) to support dynamic voltage and frequency scaling (DVFS) [9, 1, 10, 11, 7, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23]. In deep nanometer process aging, temperature and process variations affects performance of the digital circuits [30, 31]. As the IVRs are designed in the same process nodes as digital circuits, they also suffer from same variations [1, 34, 3]. In particular, variations in on-chip/on-package passives (inductance and capacitance) causes shift in the IVR's characteristics including transient response to load step and reference step [1]. The variations in the IVR's output response creates additional uncertainty for the digital circuits potentially increasing the error rates. Therefore, it is important to develop auto-tuning mechanisms for IVR to minimize the effect of IVR's variations on the voltage/timing margin or timing error rates of digital cores. The existing post-silicon auto-tuning methods for the IVRs [2, 3] involve adjusting the controller transfer function to optimize a tuning cost. However, they are not aware of the process variation in the digital logic.

In this chapter an auto-tuning method for IVR driven by the performance of the digital cores (Fig. 3.1) is presented. We propose to tune the IVR using a cost function that directly captures performance of the digital core with the objective to increase the maximum operating frequency of the digital circuit under (a) process variations in cores and passives, and (b) supply noise due to dynamic load transitions, while ensuring stable IVR operation. The proposed tuning method is based on two cost metrics to represent the system performance: (1) the accumulated sum of the delay slack of a digital core with respect to a target frequency and (2) error count during operation with a target operating frequency. To evalu-

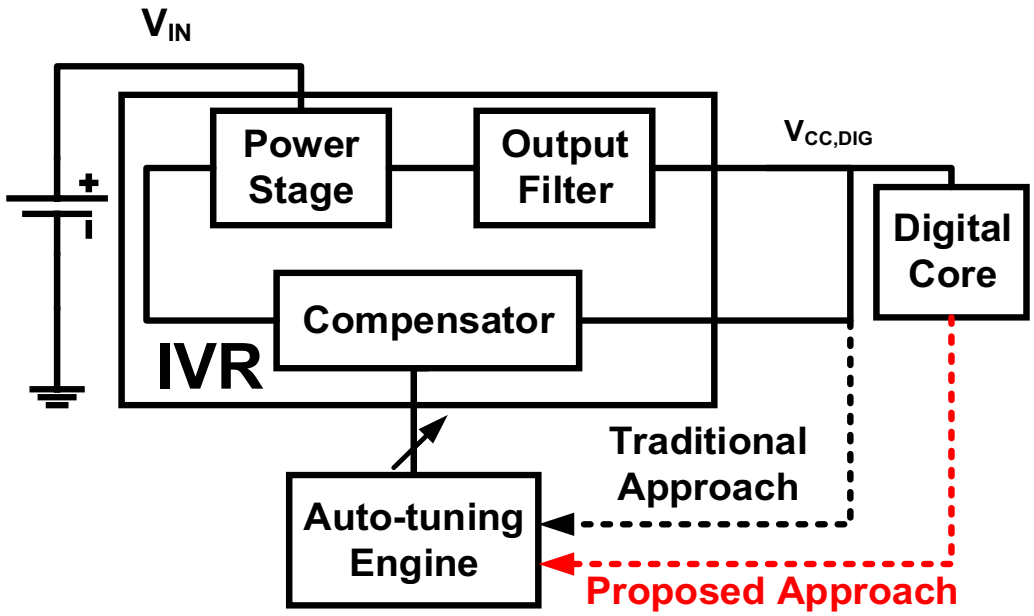


Figure 3.1: A system of an inductive IVR and digital core along with proposed auto-tuning method

ate the tuned systems, we consider a fully integrated inductive IVR (FIVR) with a voltage mode digital PWM control driving a digital core.

3.1 Motivation

We performed measurements on an 130nm test-chip [1] with IVR powering an AES (Advanced Encryption Standard) encryption engine to experimentally characterize role of FIVR controller's coefficients on run-time timing error. The FIVR power stage uses two consecutive bondwires with a total of 11.6nH inductance, 3.2nF MIM capacitance and 125MHz switching frequency. The direct form digital controller is sampled at 250MHz frequency and the compensator coefficients are reduced to 6-bits. A 128-b AES engine is driven by the IVR and is used as a digital load to the IVR. For a given clock frequency, multiple AES encryption events are executed. Depending on the supply noise and the target clock frequency, there can be timing violations in the AES causing an incorrect encryption. The AES outputs are compared with the golden responses to find out the error rate and this ex-

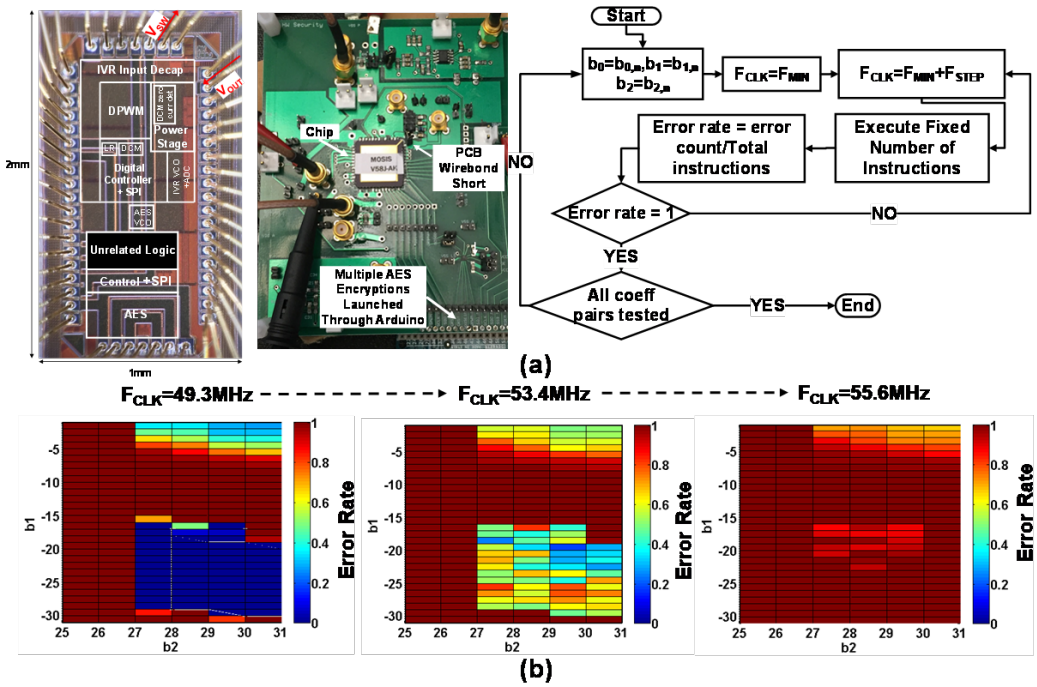


Figure 3.2: (a) Chip micrograph, test PCB, and measurement procedure for the 130nm test-chip (b) Measured error rate across different compensator coefficients (b1 and b2) for increasing clock frequencies (VCC,AES=0.8V)

periment is repeated for different IVR coefficients and increasing clock frequency (F_{CLK}). For a F_{CLK} of 49.5MHz (0.8V FIVR output) multiple FIVR coefficients yield a zero error rate (Fig. 3.2). As the frequency is increased, the error rate starts to increase for the aforementioned coefficients and becomes dependent on the FIVR coefficients. The measurement shows that timing error rate for a target frequency is dependent on IVR coefficients.

3.2 Proposed Tuning Methodology

We propose two tuning costs to quantify the performance of a digital system. We show that using these cost metrics enable (1) obtaining a stable response at DC loads and fast recovery from transient droop, and (2) optimize performance of the system under process variation of the digital core. The tuning engine generates different coefficients for different operating conditions of the digital load.

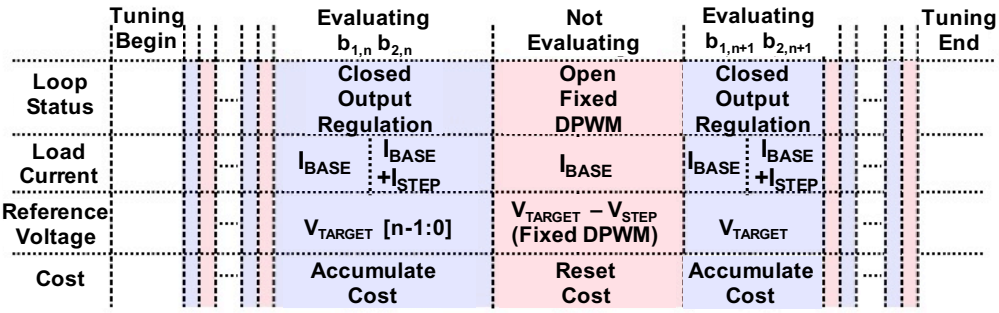


Figure 3.3: IVR control flow during the proposed auto-tuning process

3.2.1 Control Flow of Tuning Process

During the tuning process, the system performance is measured for different compensator coefficients. The control flow of the tuning algorithm as well as each evaluation period is elaborated in Fig. 3.3. Before each coefficient is evaluated, the control loop is opened (power stage driven by a fixed duty cycle) to ensure same initial condition. Unlike [1], before starting evaluation, the control loop is closed with a reference voltage lower than the target voltage. The difference between reference and the target voltage determine the reference step (V_{STEP}). After the loop is closed, the output stability at a base load current (I_{BASE}) is observed and at the middle of the evaluation cycle, a load step (I_{STEP}) is applied. The second half of the evaluation period observes the stability at current I_{STEP}+I_{BASE}. During actual runtime, the total numbers of transient droop events depend on the underlying application. However, as every coefficient goes through the same evaluation period, fairness is ensured during the optimization. The response of the control loop to the transient events during the evaluation period is mapped to the system performance and is captured in the quantified performance. We used two different costs for quantizing the system performance and are discussed in the following section.

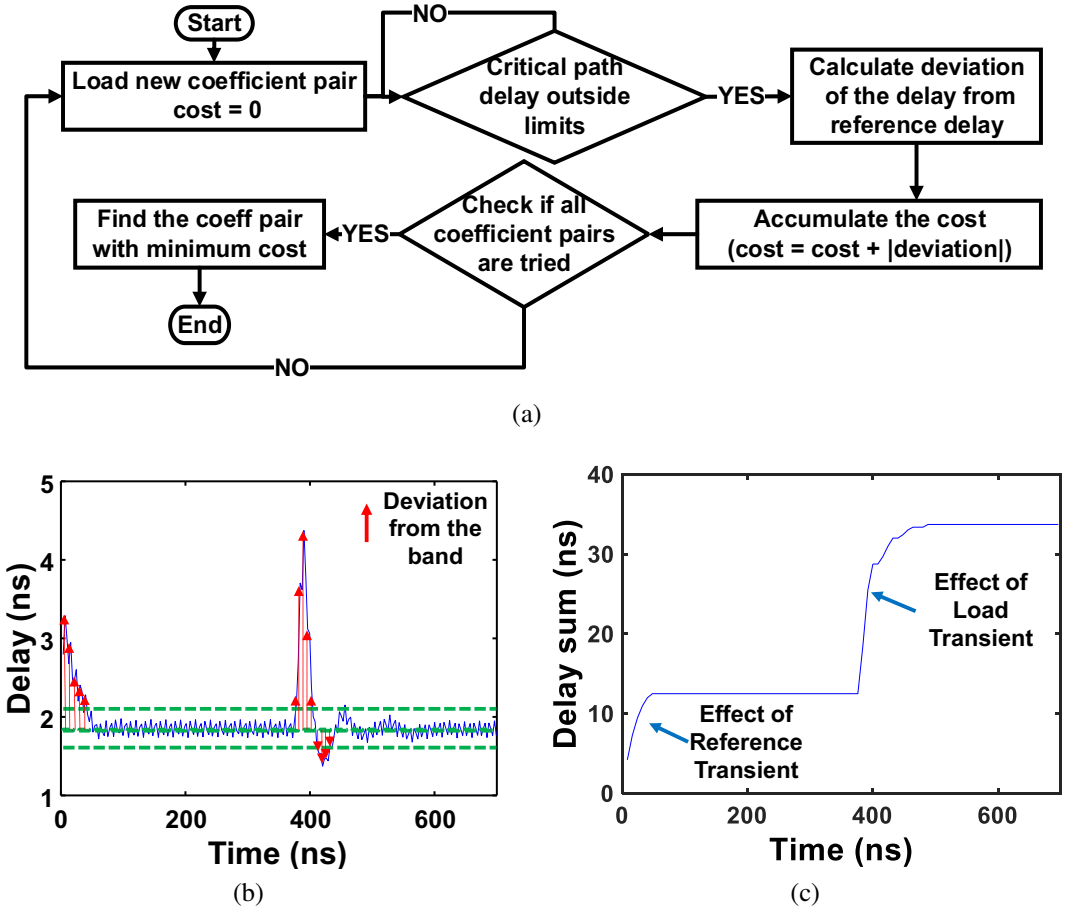


Figure 3.4: (a) Control flow of the proposed delay-sum based cost (b) An example delay response and (c) cost profile for the corresponding delay response

3.2.2 Delay-sum Based Tuning

Instead of accumulating the absolute IVR output error as part of tuning cost in [1], we propose to accumulate the absolute delay slack between the critical path delay of the combinational logic and the target clock period. Using a delay-slack based cost captures the same effect as the IVR output errors and ensures rejection of unstable coefficients, but also helps fine-tuning the response for process variation in the core. The outcome of minimizing the proposed delay-slack based cost selects a set of coefficients where across the evaluation period, the logic delay stays closest to the target delay. Fig. 3.4a shows the control flow for delay based tuning. The effect of the initial reference transient and load transient as well as the effect of IVR steady state output voltage ripple is captured in the delay of

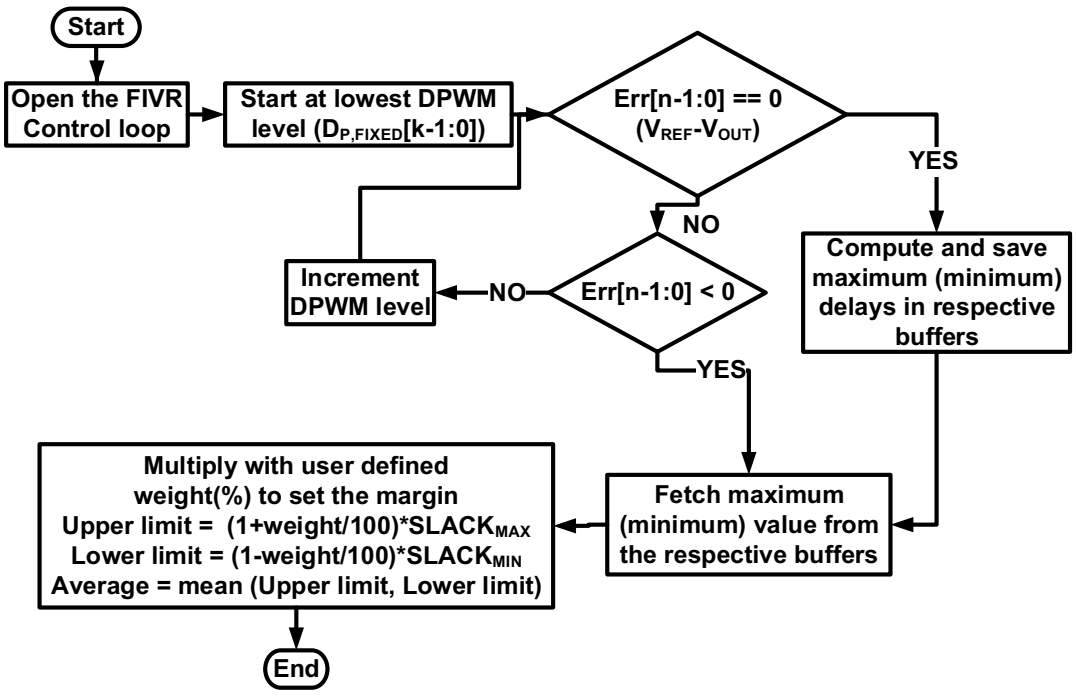


Figure 3.5: The proposed open loop test for characterizing the variation in critical path delay for steady state IVR variation

the logic. Although transient responses can be tuned by changing the compensator coefficients, the steady state ripple at the IVR output does not get affected by coefficient tuning. We propose to use a band around the delay slack and any digitized slack value within that band is neglected during accumulation to eliminate the effect of supply ripple. The band is determined using the test shown in Fig. 3.5. To determine the bands, the IVR control loop is opened, DPWM is driven by a fixed input and the digitized error is observed (Fig. 3.5). DPWM resolution is generally set higher than ADC resolution to avoid limit cycling, so for multiple $D_{P,FIXED}$ values Err_{DIG} will be zero. When the control loop is closed and the output is regulated, the $D_{P,FIXED}$ can settle to any of these values at a steady state condition. The minimum and the maximum digitized slack is calculated for these $D_{P,FIXED}$ levels, which represent the variation in the logic delay at a steady state of the IVR and cannot be tuned by changing coefficients. To account for this effect, the maximum and the minimum delay values should be multiplied with a factor to account for small fluctuations at the output voltage. The reference clock frequency can be chosen from the mean of all the delay sam-

ples collected during the open loop test. To integrate proposed cost into existing hardware we propose to use a tunable replica circuit (TRC) followed by a Vernier delay chain (VC), which acts as a time-to-digital (TDC) converter, to quantify the critical path delay [36]. The absolute value of the digitized delay is aggregated over the evaluation period of one coefficient and minimized across different sets of coefficients to obtain the optimum coefficients. Depending on the critical path obtained during synthesis, selectable fixed length portions of the cells are chosen to mimic the critical path [30]. A series of inverters are appended at the end to fine tune the TRC for delay tracking.

3.2.3 Error-count Based Tuning

For error tolerant systems such as Razor, the voltage margin is aggressively reduced to a point of first failure (PoFF) which allows a higher frequency of operation under a given supply voltage. However, total numbers of instructions correctly executed is dependent of the number of timing error detected by the error-detecting-latch [36, 37]. Once a timing error is detected, the instruction is replayed for multiple cycles till no further error is detected, leading to performance (throughput) loss. Hence, reducing he error rate, by reducing the supply noise variations is crucial to improve effective throughput. Fig. 3.6 illustrates the concept of error count based tuning of IVR's coefficients. If the voltage- margin is aggressively set, each time a large load transient occurs, there will be timing failure till the system recovers from the droop. The first droop (Fig. 3.6b) depends on the value of output capacitance and the ESR of the output capacitance, is mostly insensitive to the values of the compensator coefficients and hence the tuning process. However, the performance is also dependent on the droop settling time and the second droop (Fig. 3.6b) which can be tuned using IVR's control loop. Note, the delay based metric penalizes the coefficients if the logic delay is both lower and higher than the target delay, whereas the error count based metric penalizes the coefficients only if the logic delay is more than the target delay. Error-count based cost can be easily incorporated in IVRs powering digital engines with an

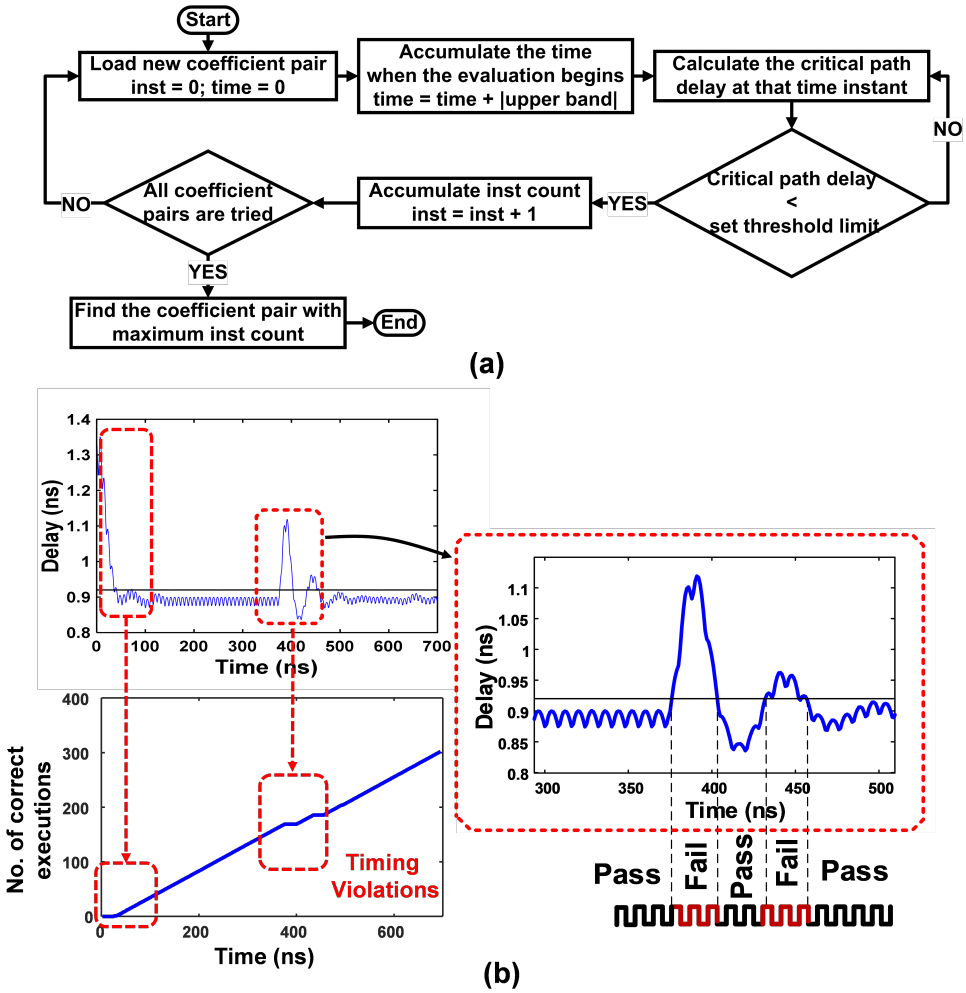


Figure 3.6: (a) Control flow of the error-count based cost (b) An example delay response for 1V VCC at nominal VT corner and corresponding cost calculation (number of correct instructions executed against time is shown

error-detection circuit. For example, the cost can be computed by accumulating the number of error events detected by a Razor latch [37] over the evaluation period. For the evaluation purpose, we set the target frequency as the upper threshold of the band found during the open loop test. This ensures that no errors are detected during the steady state operation of the IVR.

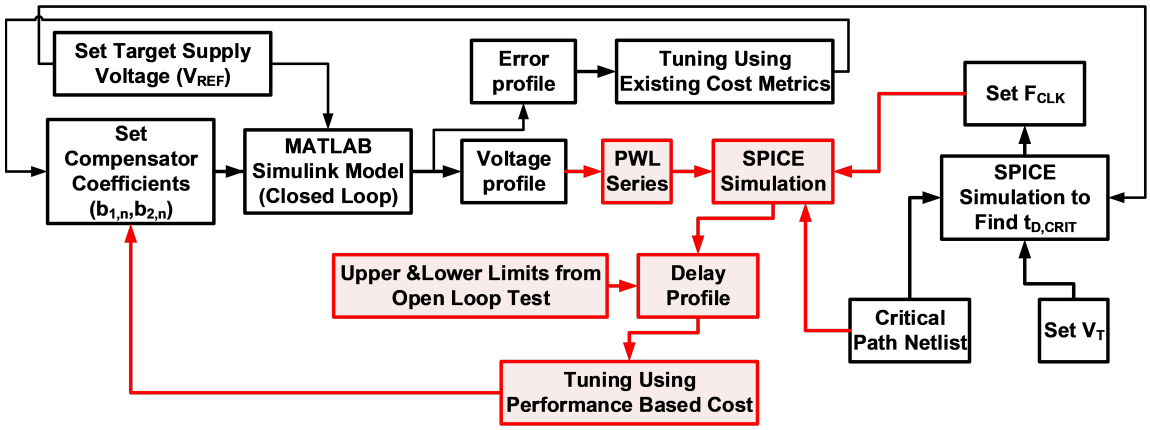


Figure 3.7: Simulation framework for the proposed performacne based tuning

3.3 Simulation Results

Fig. 3.7 shows the simulation setup for the analysis. A time-domain model of the IVR in MATLAB Simulink is used performing transient simulations. We use an IVR with 1.2V input, 6nH inductance, $50\,\mathrm{m}\Omega$ ESR, 10nF capacitance, 125MHz switching frequency with 250MHz sampling frequency. Each coefficient is represented using a 7-bit signed integer. An 8-bit ADC digitizes the difference between the reference and the output voltage. The compensator output is fed to a DPWM with 10-bit resolution (7.8ps resolution). The ADC and the compensator operate at 250MHz clock frequency whereas the DPWM operates at 125MHz. Each coefficient is evaluated for 700ns i.e. 88 IVR switching cycles. We performed experiments at two output levels: 1V and 0.7V with V_{STEP} as 0.15V and 0.1V respectively. An I_{BASE} of 10mA and an I_{STEP} of 100mA are chosen. We selected an exhaustive range for the direct form coefficients. The critical path of the digital logic is emulated as an open chain of 100 standard cell inverters in 45nm CMOS technology and simulated using SPICE. To demonstrate the advantage of the proposed tuning methodology, we use $\pm 20\%$ variation in the VT of the digital core and filter inductance (L) of the IVR at constant VCC.

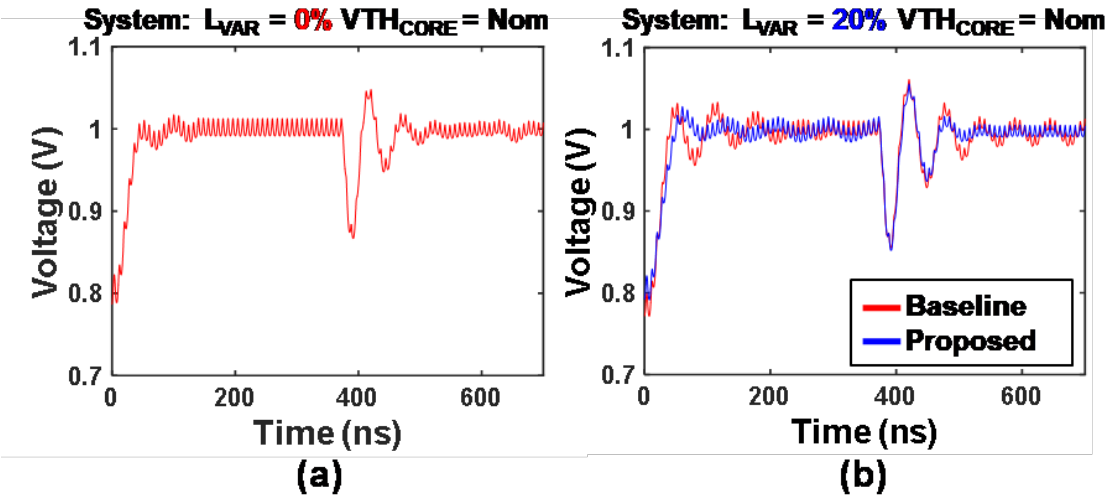


Figure 3.8: Voltage responses of IVR against passive variation, before and after tuning using existing tuning algorithms (a) At no L variation, (b) At 20% L variation

3.3.1 IVR Tuning Using Existing Algorithm

The IVR is first tuned against variation in passives. The accumulated absolute value of the IVR error samples (V_{REF} - V_{OUT}) during the evaluation period, is used as cost. The compensator coefficient pair obtained for the design with no passive and process variations using the existing auto-tuning algorithm [1], is considered as baseline coefficient pair ($C_{IVR}L_N$). Fig 3.8 shows the response of the baseline FIVR with $C_{IVR}L_N$ and the response for a FIVR with +20% variation in the inductance value using the same coefficient. After tuning, an updated coefficient is obtained ($C_{IVR}L_H$) and the response with +20% L variation improves both in terms of DC load stability as well as transient response.

3.3.2 IVR Tuning Using Proposed Algorithm

Delay-sum based cost

Fig. 3.9 illustrates tuning a baseline FIVR using delay-sum based metric can reduce effect of process variation in the digital core. First, we tune the system at nominal VT at 0.7V supply voltage (optimum coefficients $C_{SYS}L_NVT_N$). Next, we consider the digital core has moved to a high VT corner. The delay-sum based tuning results in a new coef-

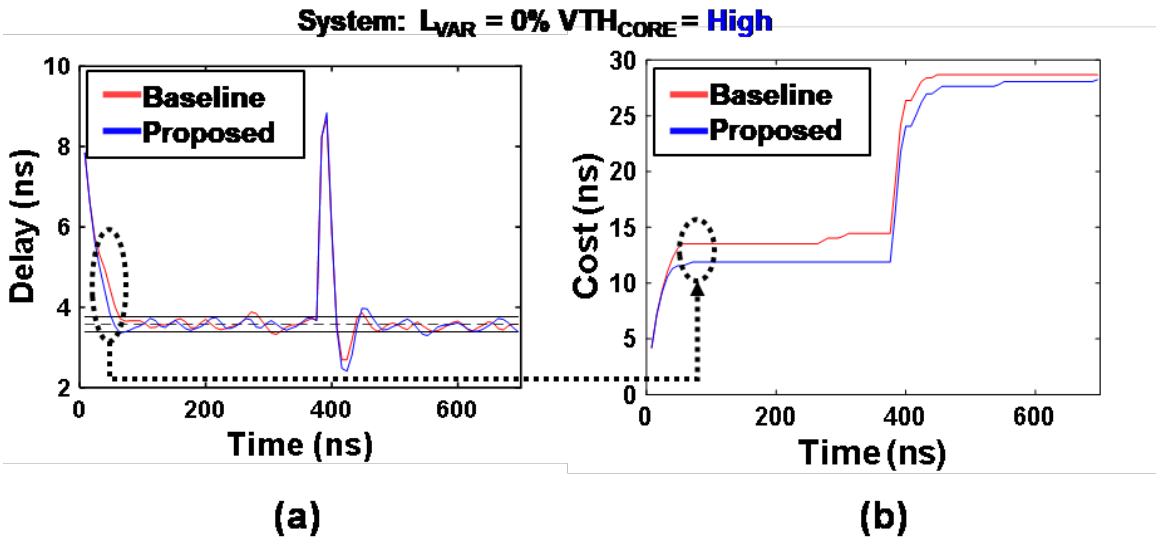


Figure 3.9: Example tuning on a system under only process variation in the digital logic (no L variation) using delay-sum based cost (a) at high VT and (b) corresponding costs against time

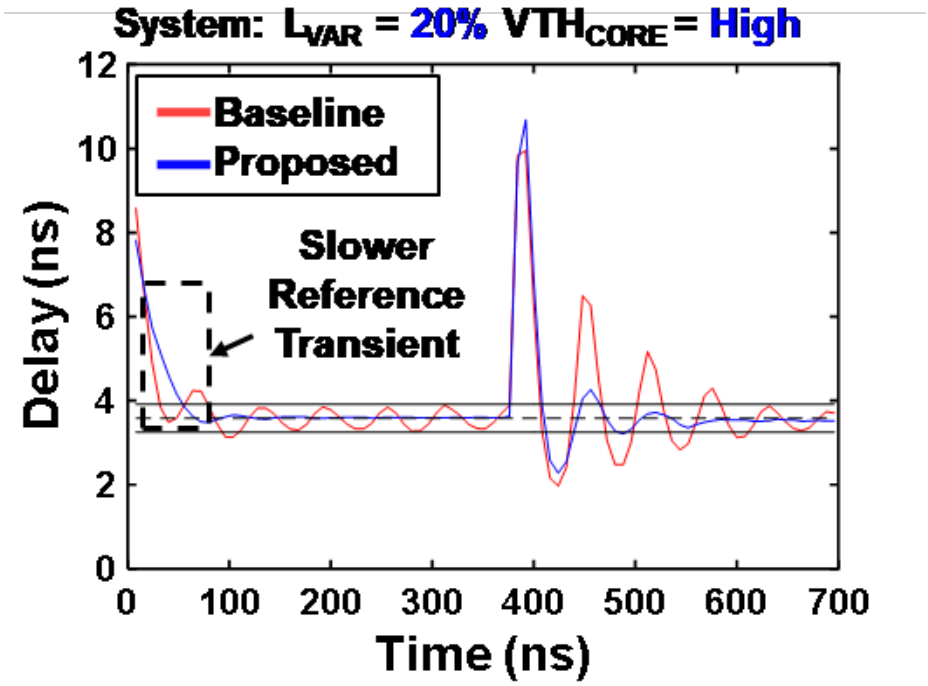


Figure 3.10: Improvement in the delay profile by using the delay-sum based cost with both process variation and passive variation at high VT and high L

ficient ($C_{SYS}L_NVT_H$). Note as L is not varying, the IVR-only tuning would have resulted in original coefficents. Fig. 3.9 shows that for the high-VT core, re-tuning the IVR with

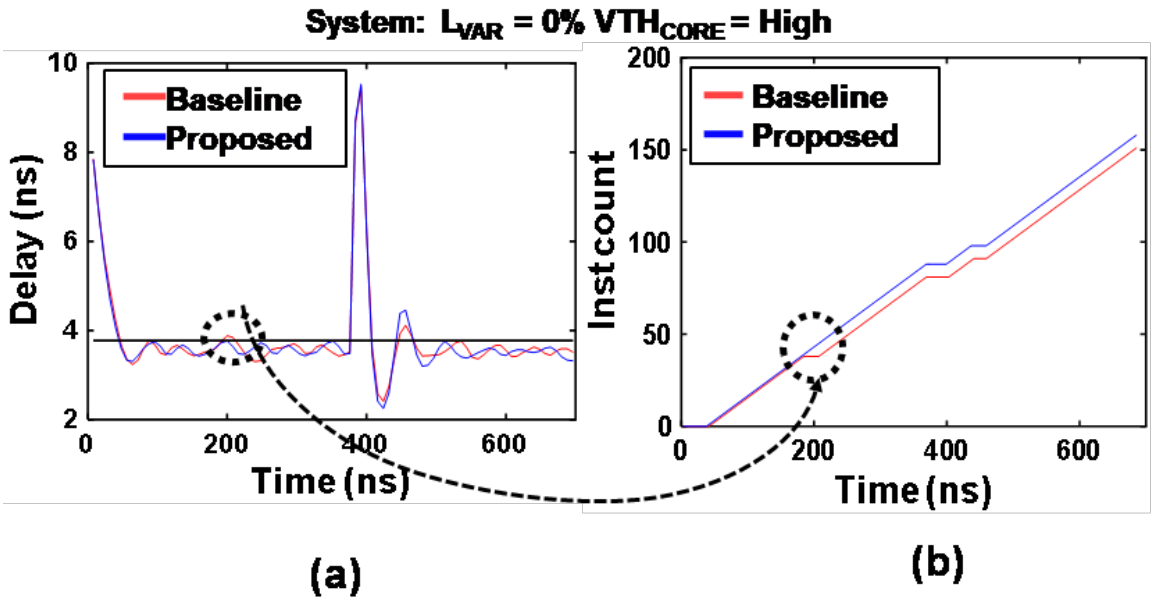


Figure 3.11: Example tuning on a system under only process variation in the digital logic (no L variation) using the error-count based cost (a) the delay profile before and after tuning and (b) the corresponding cost against time before and after tuning

$C_{SYS}L_NVT_H$ provides smaller delay variation during the reference transient (compared to the original coefficients $C_{SYS}L_NVT_N$), and causes the steady state delay variation to stay within the delay bands. The results show that delay-based tuning of IVR helps improve performance of digital core. To understand the effect of the process and passive variations we perform tuning on three individual systems, one with L only variation, one with VT only variation and last with both. Fig. 3.10 illustrates delay profiles for systems with high VT and high L. Each system is tuned to a different coefficient to reduce the delay variation especially under second droop. Note the tuned coefficients in Fig. 3.10 reduces second droop but causes slower reference transient.

Error count based cost

The same experiments were performed using an error count based cost function as illustrated in Fig. 3.11. Due to the higher voltage-delay sensitivity of the high-VT system, the coefficient $C_{SYS}L_NVT_N$ causes delay violations (at 200ns) while tuned coefficients

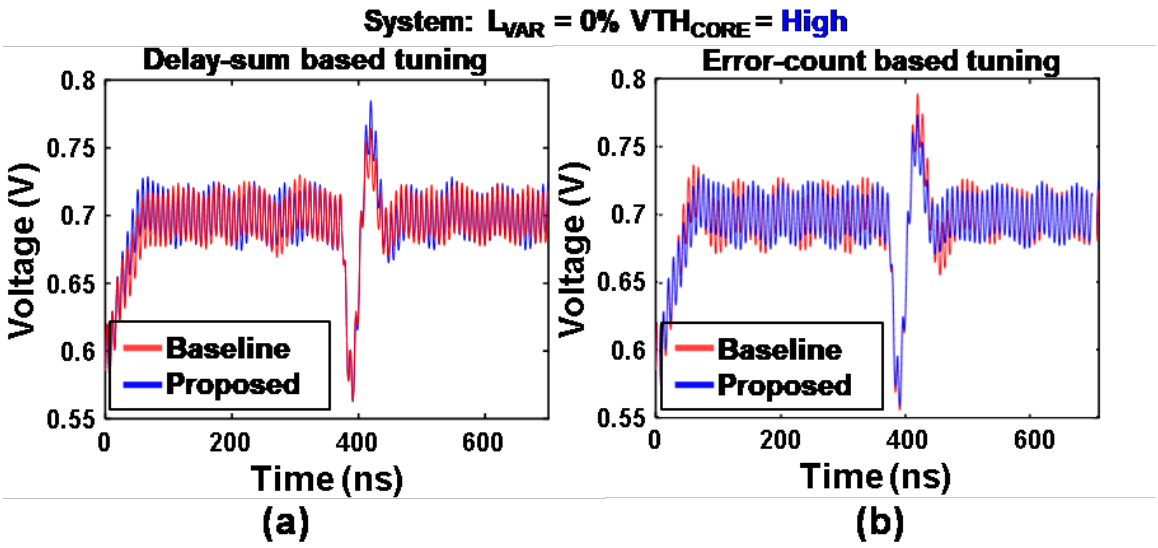


Figure 3.12: Voltage responses of IVR against process variation show improvement when using (a) Delay-sum based cost, (b) Error-count based cost

($C_{SYS}L_NVT_H$) eliminates the violation and reduces error-count based cost. Note, an error based cost is sensitive only to the duration of the delay violation, not the exact value of the delay slack. Hence, this tuning may result in more than one "optimum coefficient".

3.3.3 Impact of Performance-based IVR Tuning

In this section, we evaluate the impact of proposed IVR tuning on the system performance by estimating the error rate of the digital core at different target frequencies with tuned IVR coefficients. We apply different amounts of variations in the L, and VT_H and perform both of the proposed tuning for each system to obtain the optimal coefficients. Next, we use the optimal coefficients for the IVR, perform transient simulation of the system by applying load transient during evaluation, and estimate the error rate of digital core versus frequency. As during evaluation, the system is tested with higher than normal transient events within a time window, we call the error rate as the stressed error rate (SER). Fig. 3.13 shows SER versus frequency for a nominal L and high VT system using coefficients for IVR-only ($C_{SYS}L_NVT_N$) and proposed tuning ($C_{SYS}L_NVT_H$). We observe that delay-sum based and error-count based tuning improves frequency by 33.98MHz (@SER of 0.08) and 55.86MHz

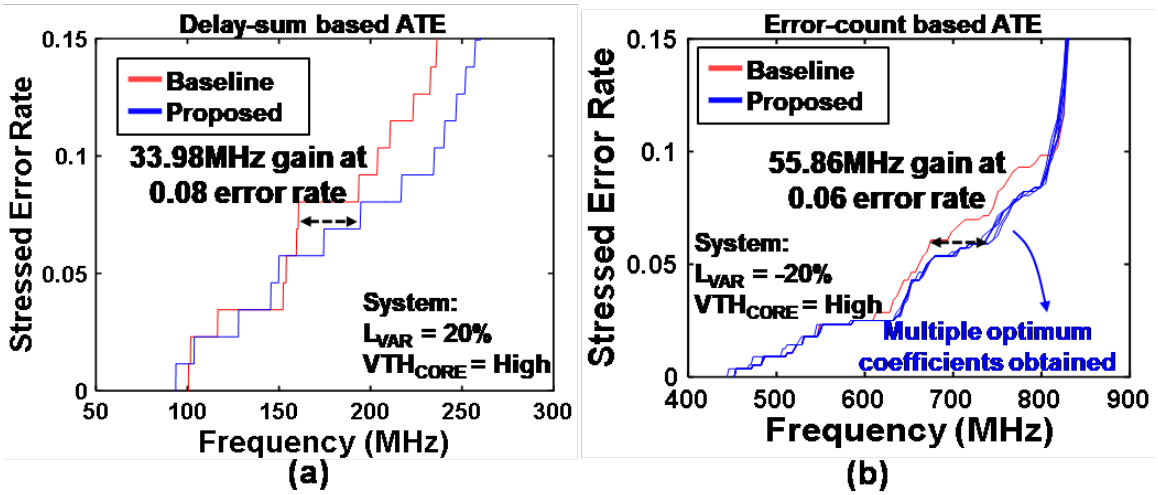


Figure 3.13: Maximum frequency gain for a given stressed (SER) rate achieved using (a) delay-based sum cost (b) error count based cost

(@SER=0.06) respectively. Table 3.1 shows the percentage improvement in frequency for SER of 0.1. We observe that improvement in frequency for a given error rate is higher for higher VT at lower supply voltage. This is attributed due to higher sensitivity of the delay to supply voltage noise which brings out the advantage of performance based tuning.

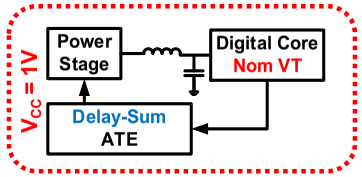
3.4 Summary

A performance based auto-tuning algorithm to tune a system of an IVR driving a digital core is presented in this chapter. We demonstrate that performance-based IVR tuning ensures a stable response with fast recovery from transient events under variations in the passives. More importantly, in the proposed approach, by tuning the IVR coefficients we can enhance the digital system performance considering process variation in the digital core and in the passives, which is beyond the capability of the existing IVR tuning methods. In conclusion, we show that the tuning of any IVR should be performed using quantifiable performance of the entire system instead of only using the performance of the IVR.

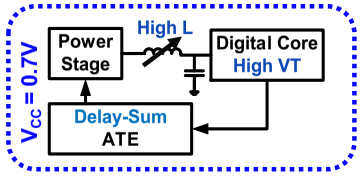
Table 3.1: Improvement in frequency for given error rate with proposed tuning

Delay-sum Based Tuning											
V_{CC}	L_{VAR} (%)	VT_{VAR}	Error Rate	Frequency Gain		V_{CC}	L_{VAR} (%)	VT_{VAR}	Error Rate	Frequency Gain	
				%	MHz					%	MHz
1	0	low	0	0	0	0.7	0	low	0.064	1.96	16.87
1	**0**	**nom**	**0**	**0**	**0**	0.7	0	nom	0.1	2.31	12.63
1	0	high	0.094	1.063	9.083	0.7	0	high	0.1	9.45	26.36
1	20	low	0.006	2.102	29.67	0.7	20	low	0.064	4.71	40.37
1	20	nom	0.012	2.457	27.57	0.7	20	nom	0.1	7.21	39.33
1	20	high	0.007	3.069	26.09	**0.7**	**20**	**high**	**0.08**	**12.18**	**33.98**
1	-20	low	0.088	1.42	2.01	0.7	-20	low	0.051	2.23	18.99
1	-20	nom	0.025	0.82	9.24	0.7	-20	nom	0.061	3.81	20.52
1	-20	high	0.061	1.32	11.25	0.7	-20	high	0.045	2.53	6.91

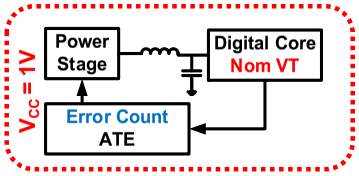
Error Count Based Tuning											
V_{CC}	L_{VAR} (%)	VT_{VAR}	Error Rate	Frequency Gain		V_{CC}	L_{VAR} (%)	VT_{VAR}	Error Rate	Frequency Gain	
				%	MHz					%	MHz
1	0	low	0.093	0.75	10.58	0.7	0	low	0.056	2.71	23.33
1	**0**	**nom**	**0**	**0**	**0**	0.7	0	nom	0.068	3.73	20.23
1	0	high	0	0	0	0.7	0	high	0.022	2.29	6.42
1	20	low	0.026	1.53	21.66	0.7	20	low	0.002	3.09	26.46
1	20	nom	0.028	1.66	18.61	0.7	20	nom	0.081	4.37	23.81
1	20	high	0.076	2.03	17.23	0.7	20	high	0.011	2.78	7.76
1	-20	low	0.061	3.85	54.29	0.7	-20	low	0.056	3.58	30.57
1	-20	nom	0.062	5.48	61.6	0.7	-20	nom	0.064	6.30	33.90
1	**-20**	**high**	**0.06**	**6.57**	**55.86**	0.7	-20	high	0.1	2.71	7.39



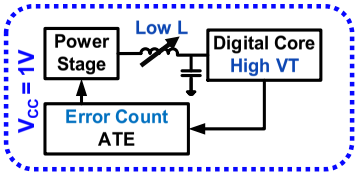
Baseline System
(Nom VT, Nom L @ 1V IVR output)



System Under Variation
(High VT, High L @ 0.7V IVR output)



Baseline System
(Nom VT, Nom L @ 1V IVR output)



System Under Variation
(High VT, Low L @ 1V IVR output)

CHAPTER 4
AUTOTUNING OF IVR USING ON-CHIP DELAY SENSOR TO TOLERATE PROCESS AND PASSIVE VARIATIONS

The analysis presented in Chapter 3 shows that performance of a digital core can indeed be improved when tuning cost includes performance metrics of digital core. However, the simulation based analysis can either underestimate or overestimate the improvement due to inherent mismatch and variations in models. In this chapter the performance-based tuning discussed in Chapter 3 is demonstration using a 130nm CMOS test-chip. The design includes a fully integrated inductive IVR (FIVR) with wirebond inductor, an on-die capacitor and a voltage mode all-digital Pulse Width Modulation (PWM) control. The IVR drives a 128-bit Advanced Encryption System (AES) engine operating at $\sim$80MHz. An on-chip Vernier Delay Line (VDL) based sensor is used to compute the delay-slack for tuning metric. An on-chip digital tuning engine generates controllable load/reference transients, computes the tuning metric using the delay-sensor, and selects the optimal coefficients of the IVR's PID controller to minimize the tuning cost.

4.1 System Architecture

In this section, we present the detailed system architecture of performance-based autotuned IVR, and tuning methodology/circuits.

4.1.1 Overall System Implementation

Fig. 4.1 illustrates the detailed architecture of the inductive IVR with the proposed performance based tuning engine. The IVR architecture has been adopted from [1]. The IVR power stage uses wirebond inductors and an on-chip MIM capacitor has been implemented.

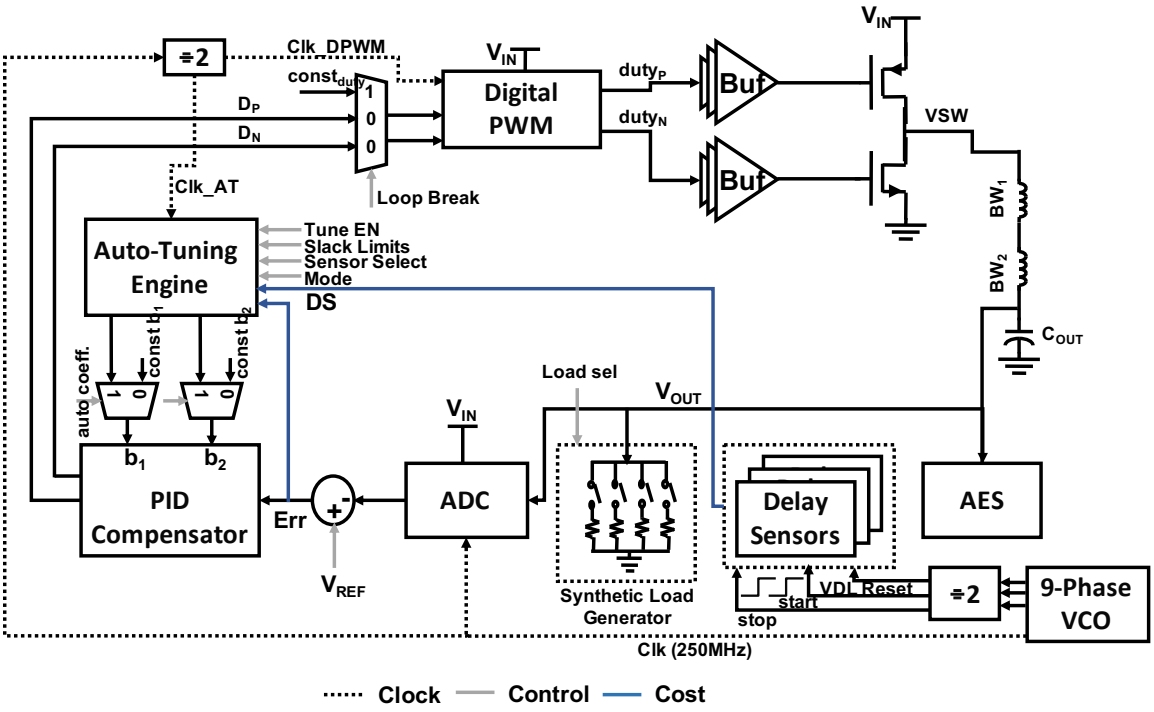


Figure 4.1: Detailed system architecture of the IVR with auto-tuning algorithm

For the digital compenstaor design a type-III compensator with two zeros is implemented as shown below:

$$G_{COMP}(z) = \frac{\frac{b_0[n-1:0]}{2^{k0}} + \frac{b_1[n-1:0]}{2^{k1}} z^{-1} + \frac{b_2[n-1:0]}{2^{k2}} z^{-2}}{1 - z^{-1}} \tag{4.1}$$

where b_0, b_1 and b_2 are each n-bit digital words and are shifted by k_0, k_1 and k_2 bits assuming a fixed-point arithmetic calculation. The all-digital compensator is fully synthesized in 130nm CMOS.

For digitizing the output voltage, a delay line ADC synchronized with the compensator clock is used. The gate signals for the power stage are generated by feeding the compensator output to a delay locked loop (DLL) based DPWM engine. To improve the loop bandwidth which dictates the response speed from transient, phase shifting of sampling clocks [14] as well as reduced precision multi-sampling [1] have been used. The multi-sampling is achieved by distributing a clock from a 9-stage voltage controlled oscil-

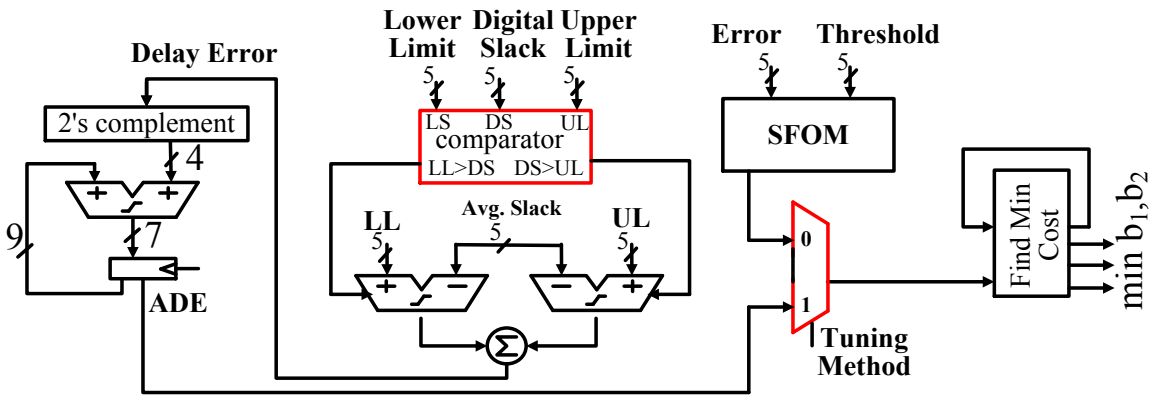


Figure 4.2: Hardware implementation of the proposed auto-tuning engine (Fig. 3.4a)

lator (VCO) to the ADC and the controller, whereas the slower DPWM clock is derived by dividing the compensator clock to ensure synchronous operation between controller and DPWM engine.

4.1.2 Hardware implementation for proposed tuning

The simplicity of the proposed tuning algorithm Fig. 3.4 allows a light and fast tuning engine operating at F_{SW} , and removes requirement for storing any digital slack samples. The use of saturated adders approximates accumulated slack for unstable/slow responses, and computes the digital slack accurately for near-optimal responses like [1]. Fig. 4.2 illustrates the hardware implementation of the proposed lightweight tuning algorithm.

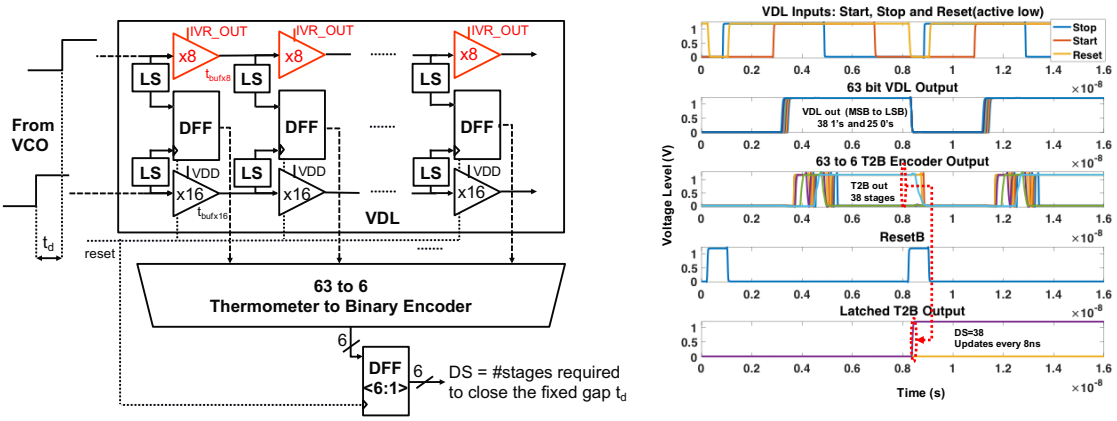
To achieve sub-gate-delay resolution, vernier configuration is used to implement the delay sensors [45, 46]. We implemented the delay sensors as a high-resolution Vernier Delay Line (VDL) with two VCO signals as inputs with a fixed $t_d = 2ns$ delay between them and the output is piped to a 63 to 6 bit thermometer to binary (T2B) encoder to obtain the digitized slack value. The resulting digitized slack value corresponds to number of VDL stages required to close the 2ns gap among the VCO signals. The relation between

the digitized slack (DS), sensor frequency (F_{sensor}) can be explained by the equations

$$F_{\text{sensor}} = \frac{1}{t_{\text{res}}} = \frac{1}{t_{\text{bufx8}} - t_{\text{bufx16}}} \tag{4.2}$$

$$DS = t_{\text{d}} * F_{\text{sensor}} \tag{4.3}$$

The implemented VDL has 63 delay stages with the resulting T2B encoder having a 6-bit output. Fig. 4.3a illustrates the designed delay sensors. For simplicity, to incorporate the effect of transient events in the cost computation, the slower of the two buffers $(bufx8)$ in the delay cells are powered by the IVR output whereas the other delay line has a constant 1.2V DC supply shared with the controller. Since the IVR output can be at different voltage levels, a level shifter is included before each of the flops in the delay stage to have appropriate voltage levels for the correct flop operation. The level shifters are included before both the flop signals to ensure equal delay is maintained in the flop signals. To avoid delay sensor output values from being saturated, the VDL flops are reset once in every clock cycle. To operate synchronously with the auto-tuning engine, the auto-tuning engine clock, fixed delay VCO signals, and the reset signal for the VDL flops are all derived from the



(a) (b)

Figure 4.3: (a) Delay sensor design for the proposed tuning cost (b) Timing diagram for the delay sensor

same 9 stage VCO. The output of the T2B encoder is latched with the reset signal for the VDL flops and is sent to the tuning engine as the digitized slack (DS). Fig. 4.3b illustrates the timing diagram for the delay sensor operation.

4.2 Measurement Results

Fig. 4.4 illustrates the measurement setup for the designed test chip. The proposed auto tuning engine for IVR is demonstrated in a test chip designed in 130nm process and 52 pin CLCC package. The inductor for the power stage is formed by two bondwires in the package shorted externally through a PCB trace, providing an effective inductance of 11.6 nH [47]. The output capacitance is implemented as a 3.2nF MIM capacitor. The power stage operates at 125 MHz and is capable of converting 1.25V input supply to 0.5 - 1V output range with the ADC resolution of ∼40mV. The 5-bit delay line ADC digitizes the

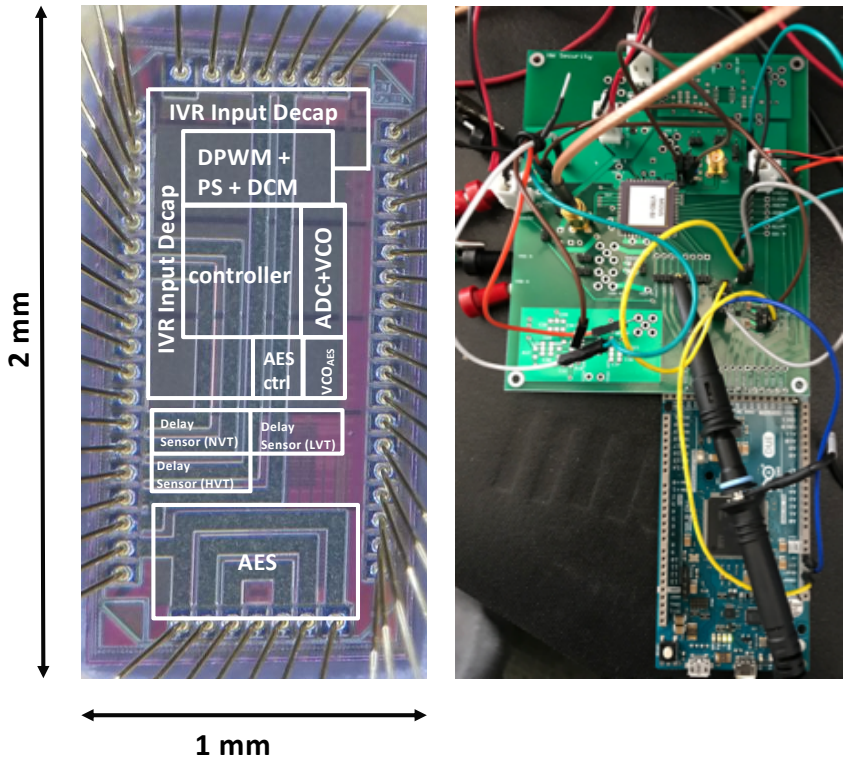


Figure 4.4: Chip micro-graph with approximate estimation (not to scale) of functional blocks and measurement setup for the designed 130nm test chip

	V_{IN}	1.25V
	V_{OUT}	0.5V-1V
	$I_{O,MAX}$	100mA
	ΔI_O	65mA
	L	11.8nH
	C_{OUT}	3.2nF (MIM)
	C_{DECAP}	1.9nF (MOS)
	F_{SW}	125MHz
Area (mm^2)	ADC	0.022
	VCO	0.016
	Top level controller (SPI+Compensator+Auto-tuning)	0.124
	DPWM	0.057
	Power Stage (Power FETs+Drivers+DCM+RTA+Load Gen.)	0.043
	IVR Input Decap	0.39
	Output Cap	0.667
	Delay Sensor (VDL+T2B)	0.024/sensor

Table 4.1: Designed IVR Specifications

output voltage. The compensator output is fed to a DPWM with 6-bit resolution. The ADC and the compensator operate at 250MHz clock frequency whereas the DPWM operates at the switching frequency of the power stage (125MHz). The output characteristics, ADC, control loop and the delay limits are all characterized by operating the power stage in the open loop condition with fixed duty cycle. Selection of ADC resolution lower than the DPWM resolution helps reducing the limit cycling.

The auto-tuning process cycles through an exhaustive range for the direct form coefficients that span across

$$b_0 = 1; -\frac{31}{16} \leq b_1 \leq 2; -\frac{31}{16} \leq b_2 \leq 2 \tag{4.4}$$

This range of coefficients translates to the following range for a parallel form conventional PID implementation.

$$-\frac{42}{16} \leq K_i \leq 4; -5 \leq K_p \leq \frac{93}{16}; -\frac{31}{16} \leq K_d \leq 2 \tag{4.5}$$

For the given range of coefficients, the frequency domain parameters such as phase margin (Φ) and bandwidth (BW) are represented by eq. 4.6.

$$-170^{\text{o}} \leq \Phi \leq 170^{\text{o}}; 5.5MHz \leq BW \leq 54.8MHz \tag{4.6}$$

It should be noted that the frequency domain analysis is obtained from simulation. There might be slight mismatch between estimated (from datasheet) electrical parameters of bondwire inductor and MIM cap for simulation when comparing with actual testchip. Thus, it will be difficult to correlate the gains with frequency domain analysis unless on-chip circuitry to perform those measurements accurately is used.

Since the main purpose of this work is to demonstrate scope for improving performance of digital cores integrated with an IVR by using delay of digital circuits as a cost metric to tune IVR loop, optimizing the design for high efficiency and high bandwidth were secondary priorities when designing the test-chip. However, a minimal frequency domain analysis is performed by reading out the coefficients from measurements results to verify that the designed and the tuned coefficients converge to stable response ($\Phi > 35^o$ and $BW > 35MHz$).

We use delay sensors designed in low-V_t (LVT) and nominal-V_t (NVT) devices to emulate process variations. Fig. 4.5 show measured sensor output (delay slack) of the two sensors, LVT and NVT along with the ADC in the operating range of 700-910mV. As expected, a higher output voltage reduces the delay of VDL, and hence, increases the measured slack (sensor output). It can be observed that the sensor outputs change linearly with the output voltage and have a wider range than the ADC. Thus, we can use the delay sensor outputs over the digitized voltage error to characterize transient supply variations. Moreover, we see that same voltage variation will result in different outputs from the two delay sensors. Hence, using the delay sensor output for tuning inherently includes the effect of process shifts on supply sensitivity.

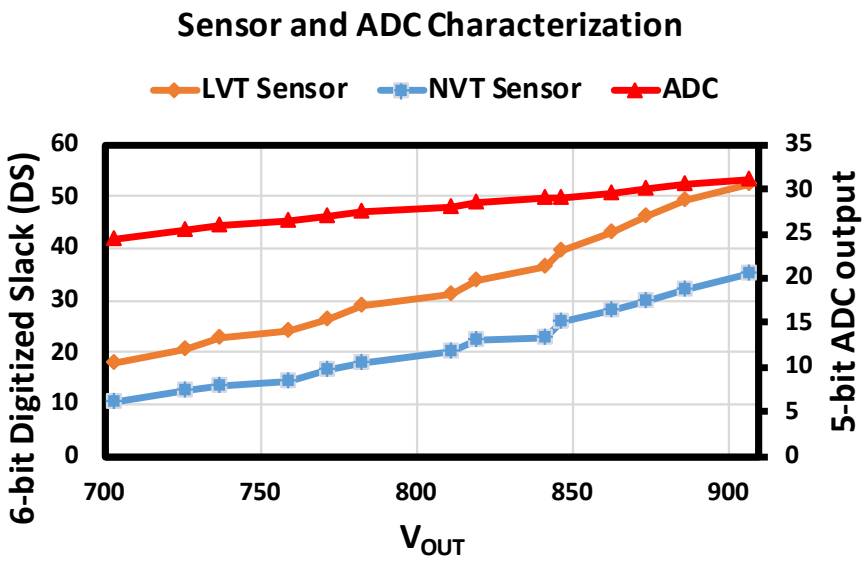


Figure 4.5: ADC and delay sensor characterization

The impact of auto-tuning on performance is measured in two ways. First, we consider the delay sensors themselves as load to understand whether performance-based tuning can improve the average frequency of the sensors, compared to using un-tuned coefficients. Second, a 128-bit parallel AES engine is integrated in the test-chip to measure performance improvement in an actual digital logic. The measurement of the AES engines shows that tuning coefficients using the delay sensor output helps to improve performance of digital engines under process and supply variations.

4.2.1 Auto-tuning Process

Fig. 4.6 illustrates the IVR output when the auto-tuning process is enabled using an external controller. The controller disengages the feedback loop to find the optimum coefficients and then re-engages the loop once the new coefficients are loaded. The tuning engine operates at power stage frequency (125MHz) and for the exhaustive search across the coefficients the total tuning time is $\sim$16.7ms. This time can be reduced by narrowing the search space as the maximum and minimum coefficient limits are user programmable.

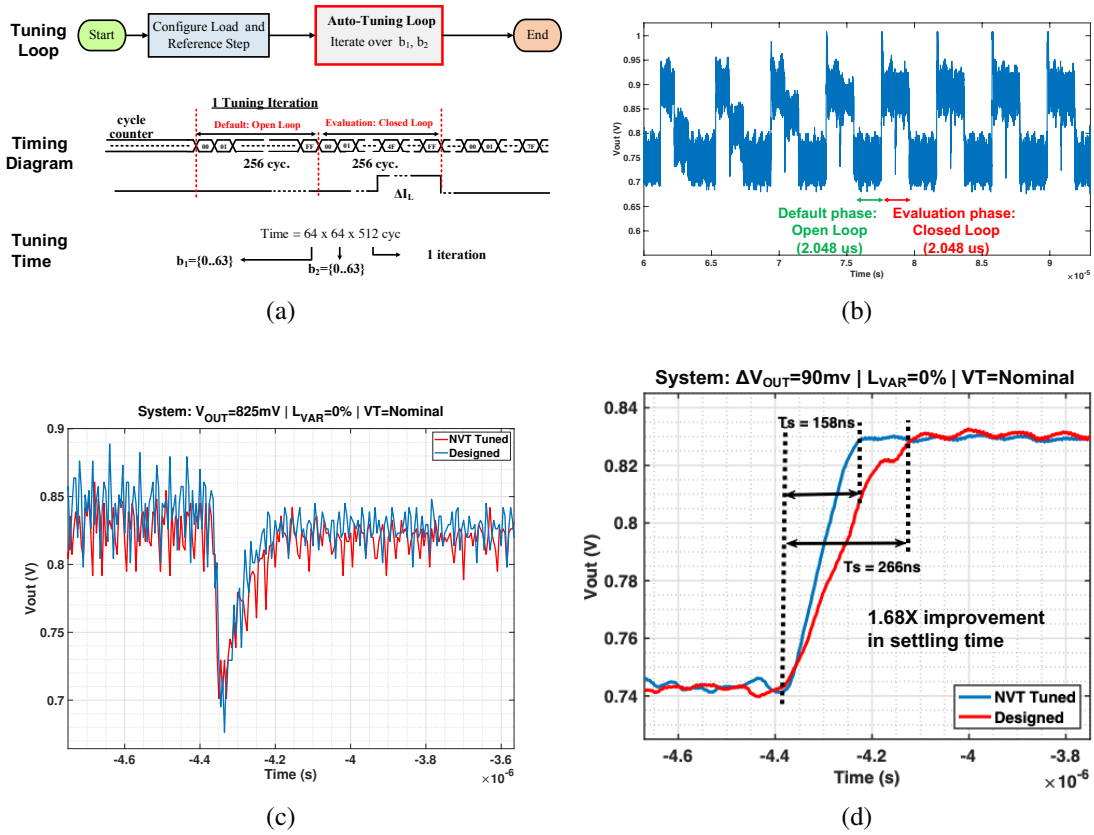


Figure 4.6: (a)Timing diagram for the tuning process, (b)Zoomed in figure of IVR output during the auto-tuning process (c)Load transient response for designed coefficient and the tuned coefficient obtained from the proposed tuning algorithm (d) Reference transient response (band-limited) for designed coefficient and the tuned coefficient obtained from the proposed tuning algorithm

Tuning the Designed coefficients

The IVR is first used with the designed coefficients obtained during modelling and simulation of the test chip from the transfer function of the control loop. This is considered as designed coefficient pair. Fig. 4.6c and fig. 4.6d shows the response of the IVR to a load transient and reference transient respectively with aforementioned designed coefficient and coefficient obtained from the proposed tuning algorithm using the NVT sensor for cost computation. The fast load transients (78mA/100ps) are generated using on-chip synthetic load generator and the reference transients are generated by changing the digital reference word from the external controller. The measurement results show that the tuned coefficients

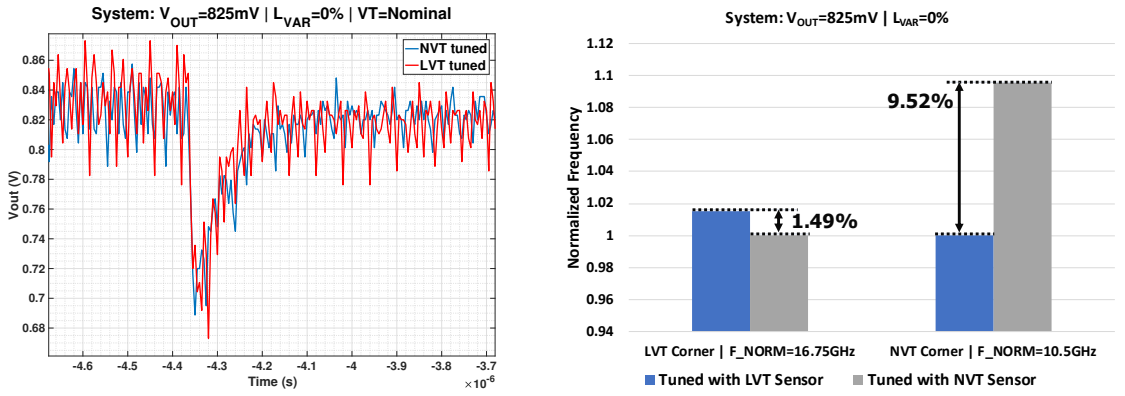
yield up to 1.68× better reference step response and a comparable load step response than the designed coefficients. The tuning results are characterized by measuring the frequency of the NVT sensor. The NVT sensor frequency measured using the coefficients generated by delay sensor based auto-tuning is 2% greater than the frequency measured using the designed coefficient pair.

Tuning against multiple sources of variations

The proposed delay based tuning algorithm tunes the feedback loop for the coupled effects of all the multiple sources of variations (passives, process, temperature, clock jitter, harmonics, limit cycling, package resonance, etc.). This is possible since the combined effects of these variations are mapped to the IVR output response which supplies one of the delay chains in the sensors. Thus effects of all these variations can be accurately captured to the delay profile for cost estimation and minimization during the tuning process. The proposed tuning cannot identify the sources of variations. Hence we are testing for the only controllable sources of variations: a) passives (by adding additional inductors in series) and b) process (by using LVT and NVT cells for delay sensors).

Tuning against only process variations

Let us consider that the baseline condition has no variations and thus the coefficient obtained using the NVT sensor (Fig. 4.6c) is the baseline now. Now assume the process shifted to LVT corner. We have two options: (a) use the baseline coefficients tuned with NVT sensor also in the LVT corner, and (b) re-tune the IVR considering the LVT sensor output and obtain a new coefficients. Note as L is not varying, the traditional tuning [1] would have resulted in original coefficients. Fig. 4.7a illustrates the load transient response for the IVR for this analysis when tuned with both NVT and LVT tuned coefficients. Fig. 4.7b shows how re-tuning for process variation changes the sensor frequency. We observe that LVT sensor shows 1.49 % higher frequency when the IVR operates with coefficients



(a) (b)

Figure 4.7: For system under process only variation: (a)Load transient response for the system with coefficients tuned using different sensors (b)Delay sensor frequency improvement using the proposed tuning algorithm

auto-tuned with the LVT sensor, compared to when tuned with NVT sensor. Likewise, NVT sensor shows a 9.52% higher frequency when the IVR is re-tuned to tolerate process shifts from LVT to NVT corner.

Tuning against passive and process variations

The final configuration tested is when the system undergoes variations in both process and passives. To model passive variations we add a 6nH inductor in series between the two bondwires and connect them using PCB traces. This results in +50% inductor variation. The baseline coefficient is the one obtained from tuning with NVT sensor with no L variation. The new coefficients for the system with the extra inductance are obtained by tuning with NVT and LVT sensors. Fig. 4.8 illustrates sensor frequencies for the system under variation with maximum gain of 13.9% when the system moves from LVT corner with no passive variation to NVT corner with 50% inductor variation.

4.2.2 Impact of Process Variations on IVR Performance

The impact of process variations on performance of IVRs can be classified mainly into two categories. For analysing both the cases, consider process variations to be mapped as

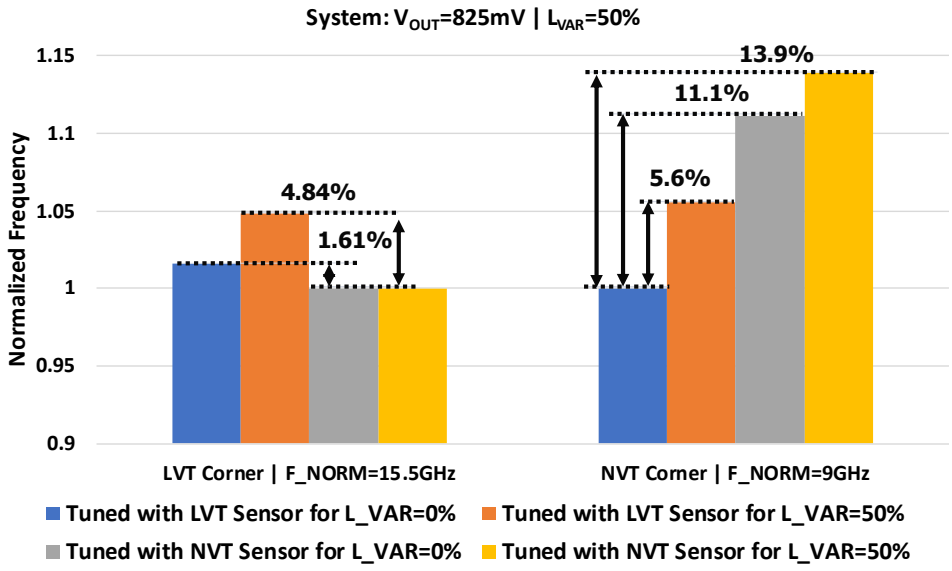


Figure 4.8: For system under NVT and 50% L variation: Delay sensor frequency improvement using the proposed tuning algorithm

change in V_{TH} that leads to change in on-resistance (R_{ON}) of FET devices.

IVR Power Stage

The frequency response of the power stage is dominated by the LC poles which are relatively insensitive to R_{ON} of power stage FETs. Thus, a change in the power stage resistance has a very limited to negligible effect on transient response which will not change the supply noise/DC shift of IVR output when compared to that of before the variations. Hence tuning wouldn't be of much use in this case. Although in this case efficiency of the IVR will go down.

IVR controller and digital core:

The change in process corner of the devices in the controller and the digital core will lead to different values for load transient events and settling times of the IVR output response when compared to before variation. This in-turn affects the supply noise/DC shift of the IVR output. Since the sensors also undergo the same variations, the effect on the IVR

output response at different corners can be accurately mapped to the delay profiles and tuning will lead to a design with minimal supply noise and DC shift.

4.2.3 Performance Improvement of the Digital Core

The performance of IVR is characterized mainly as droop and settling time. A higher supply noise or DC shift in IVR output due to variations may lead to higher droop and/or longer settling time increasing the timing errors in the digital core. Thus, there is direct impact of performance of IVR (response speed and voltage droop) on performance of digital core. Hence, proposed tuning to improve the transient performance of IVR can lead to possibility of improving performance of the digital core.

For the performance measurement, we have implemented a 128-bit parallel AES engine as a digital load which is driven by IVR output. Here performance is considered as the maximum operating frequency of the AES core without any timing errors. Thus to measure the performance improvement of the AES core, we perform 1M plain text encryptions at different AES clock frequencies. At each frequency for an target error rate (TER), number of incorrect encryptions is also calculated. The maximum operating frequency is measured as the highest clock frequency that ensures no errors (TER=0). The design only include an AES engine designed with NVT devices. The preceding measurement is performed by inducing load transients during AES operation to create transient supply noise. A higher supply noise i.e. higher droop and/or longer settling time increases the error rates. The AES error rate test is performed with a 50 % L variation. During testing we considered IVR auto-tuned for following four conditions: (i) tuning with LVT sensor and no L variation, (ii) tuning with LVT sensor and 50 % L variation, (iii) tuning with NVT sensor with no L variation, and (iv) tuning with NVT sensor and 50 % L variation. It is evident the last case corresponds to the actual measurement condition of the AES engine. Consequently, as expected, the maximum AES frequency is observed when the IVR is tuned in the same condition (i.e. case iv), as illustrated in Fig. 4.9. We have also re-tuned the system for

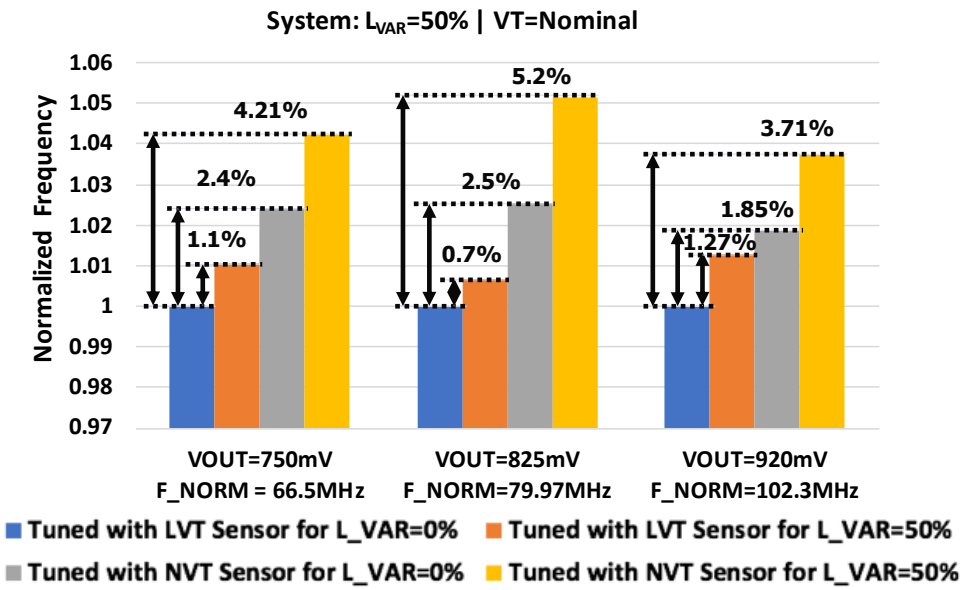


Figure 4.9: Improvement in the performance of the AES core due to the proposed tuning

different supply voltage conditions, and observed that preceding observation is valid. The maximum performance improvement observed for the AES core is upto 5.2% (4.31MHz).

4.2.4 Power Efficiency

Fig. 4.10 illustrates measured power efficiency of the designed IVR for different load currents. The measured efficiency of the designed IVR attains a peak value ∼69% at 825mV

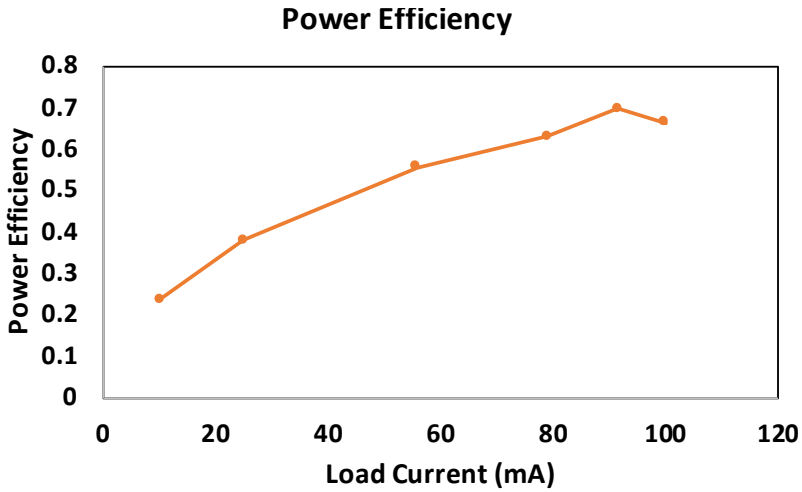


Figure 4.10: Measured power efficiency for the designed IVR system across different load current

and 91mA load current. The efficiency measurement considers losses in power stages, drivers, ADCs, sensors, and controllers. In other words, the resistive and switching losses are included in the measurement. We further observed that the efficiency is unchanged when using the coefficients obtained from the proposed auto-tuning process versus using the designed coefficients.

4.3 Summary

This chapter experimentally demonstrates a performance based auto-tuning of an inductive IVR introduced in Chapter 3 driving a digital core, an AES engine, in 130nm CMOS. The proposed tuning ensures a stable response and improves transient response under variations in the passives. More importantly, in the proposed approach, by tuning the IVR's coefficient we can enhance the digital system's performance considering process variation in the digital core and in the passives. In conclusion, we show that the tuning IVR using quantifiable performance of the entire system, instead of only using the IVR's output, helps improve system performance considering variations in transistor process and passives.

Table 4.2: Comparison With Prior Work

	[15] (2013)	**[16] (2014)**	**[10] (2011)**	**[13] (2016)**	**[9] (2012)**	**[1] (2017)**	**This Work**
Technology	130nm	22nm-Trigate	130nm	65nm	45nm	130nm	**130nm**
L (nH)	3-7 (Bondwire)	1.5 (On-Die)	2 (On-Die)	1.54x2 (On-Die)	26x4 (SMT)	11.8 (Bondwire)	**11.8 (Bondwire)**
C (nF)	9.8 (On-Die)	10 (On-Die)	5 (On-Die)	1.83 (On-Die)	23 (SMT)	3.2 (On-Die)	**3.2 (On-Die)**
$^{+}$FSW (MHz)	100	500/500	300	500/500	80	125/250	**125/250**
Controller	Analog PWM	Digital PWM	Analog PWM	Digital PWM	Analog non--linear PWM	Digital PWM	**Digital PWM**
V_{IN} (V)	1.2	1.5	1.2	2.0-2.2	1.5	1.2	**1.25**
V_{OUT} (V)	0.9	1	0.86	1.2	0.6-1.3	0.45-1.05	**0.5-1**
Peak Eff (%)	82 (300mA, 0.9V)	68 (90mA,1V)	74.4 (125mA,0.86V)	76 (400mA,1.2V)	83 (750mA,1V)	71 (50mA,0.8V)	**69 (91mA,0.825V)**
Area (mm^2)	2.25	1.5	0.56	1.13 (1.59 including decap)	0.75	0.5 (1.19 including MIM cap + decap)	**0.5 (1.23 including MIM cap + decap + 3 sensors)**
+: Switching frequency/Sampling frequency (for digital controller)							

CHAPTER 5

AGING CHALLENGES IN ON-CHIP VOLTAGE REGULATOR DESIGN

The simulation and measurement analysis presented in the previous chapters clearly emphasises the need for tuning against passive and process variations. In this chapter another major inevitable source of transistor variation is discussed that further reinforces the need for tuning of on-chip voltage regulators. Transistor aging mechanisms including bias temperature instability (BTI), hot carrier injection (HCI), time dependent dielectric breakdown (TDDB), and electromigration (EM) are becoming more prevalent with the rapid scaling of process nodes. BTI is a very common yet critical reliability concern for most nanometer integrated circuits design [48, 49, 50]. However, the performance degradation induced by BTI is generally overlooked for on chip VRs [51, 52, 53]. There are other sources of aging as well that might affect transient performance of on chip VRs. Since VRs are typically sourcing DC current (for linear regulators) or AC current (inductive regulators), they would be sensitive to hot-carrier injection (HCI) degradation as well. In linear regulators such as digital low drop out regulators (DLDOs) different devices will be stressed differently over their lifetime depending on load current patterns. This non-uniform stressing of the devices may lead to additional ripple and limit cycling oscillations at the output of the DLDO. There has been some recent work [52] which discusses mitigating NBTI/HCI based degradations in DLDO by using unidirectional shift registers to ensure uniform aging of the PFETs in the power stage. However, detailed analysis of aging effects on on-chip voltage regulators considering aging in different sections of the design as well as low overhead reliability enhancement techniques under arbitrary load conditions have not yet been completely investigated and verified via silicon measurements.

This chapter analyzes the reliability of two types of on-chip VRs, namely, DLDO and IVR, due to NBTI effects on the power stages. We present simulation-based analysis and

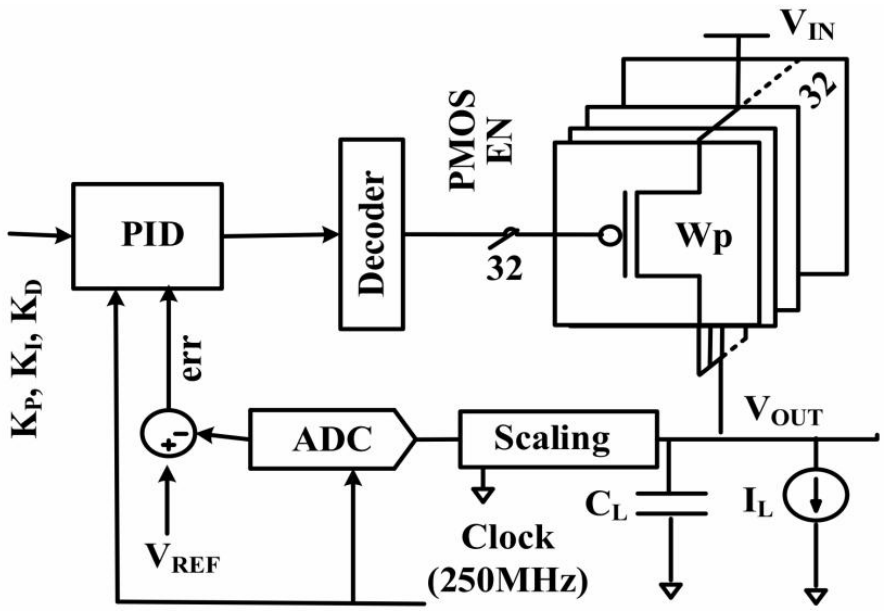


Figure 5.1: Architecture of Digital LDO

measurements from 130nm [53, 18] and 65nm [54] CMOS test-chips to characterize and compare degradation in the transient performance and power conversion efficiency of on-chip VRs due to NBTI.

5.1 Design and Modelling of On-chip VRs

5.1.1 DLDO Design and Modelling

Fig. 5.1 shows overall architecture for implemented DLDO system. DLDO power stage consists of 32 PMOS devices. The DLDO feedback loop consists of a delay line based 4-bit analog-to-digital converter (ADC) followed by a type III proportional integral derivative (PID) compensator implemented in parallel form. A decoder converts a 5-bit word from PID to a 32-bit control signals for the power stage to perform regulation. A decoupling capacitance is used at the output of power stage. The entire feedback loop runs at a fixed clock frequency (no multisampling). To understand overall control loop, we present a z-

domain model for the DLDO. The z-domain model of the ADC is represented as:

$$H_{ADC}(z) = \frac{1}{v_{LSB}} \times z^{-1} \tag{5.1}$$

where v_{LSB} is the analog voltage change for 1 LSB (least- significant-bit) difference in the digitized ADC output. The power stage (PMOS array) is compensated with a type-III (two zeros and two poles) PID controller. The z-domain transfer function is given as:

$$H_C(z) = \left[k_p + \frac{k_i}{1 - z^{-1}} + k_d(1 - z^{-1})\right] z^{-1} \tag{5.2}$$

where k_p, k_i and k_d are proportional, integral and derivative gains respectively. The compensator is implemented with fixed/reduced precision arithmetic to ensure timing constraints are met. The inputs to the compensator are 5-bit K_P, 5-bit K_I and 4-bit K_D digital words. Since the output of PID compensator is registered, a single cycle delay (z_{-1}) is incorporated in z-domain transfer function.

The output of the compensator controls the PMOS transistors in the power stage through a zero-order-hold (ZOH). The transfer function for the power stage (PFET array) in z-domain can be represented as:

$$H_P(z) = \frac{K_{DC}(1 - e^{-\omega_L/F_s})}{(z - e^{-\omega_L/F_s})} \tag{5.3}$$

$$\omega_L = \frac{1}{(R_P||R_L)C_L} \tag{5.4}$$

$$K_{DC} = I_{PMOS} \times R_P||R_L \tag{5.5}$$

where R_P and R_L are power stage and load resistance respectively (generally $R_P << R_L$), ω_L is the load pole, I_{PMOS} is current capacity of single PFET device in the array and F_s is

the sampling frequency. For steady state analysis the power stage is modelled as an effective resistance R_P. However, for transient analysis including load and line regulation, the power stage model is based on PFET current equations in linear and saturation region. In steady state, PFET array is assumed to be in linear region with a constant device resistance. Effective resistance of the power stage is determined by dropout voltage (V_{DO}) and load current (I_L) as, $R_P = {}^{V_{DO}}/_{I_L}$.

Open loop transfer function for the DLDO system can be derived with z-domain transfer functions for the power stage, ADC and the compensator. Closed-loop transfer function can be derived from open-loop transfer function as follows

$$H_{OL}(z) = H_{ADC}(z)H_C(z)H_P(z) \tag{5.6}$$

$$H_{CL}(z) = \frac{H_{OL}(z)}{1 + H_{OL}(z)} \tag{5.7}$$

It can be observed that the DC gain of the open loop transfer function will be mainly dominatied by K_{DC}, the power stage gain.

5.1.2 IVR Design and Modelling

Fig. 5.2 illustrates the architecture of the IVR system implemented. The output filter of the power stage is implemented using an inductor and a capacitance. The feedback loop of IVR is very similar to that of the DLDO system containing an ADC, PID controller (implemented in direct form) and a digital pulse width modulation (DPWM) block. The compensator output (digital control word) is fed to a delay locked loop (DLL)-based DPWM engine, generating gate signals with a duty cycle based on control word. The implemented system incorporates multisampling, i.e. the ADC samples the output at twice the rate at which power stage is operating. ADC and compensator operate at twice the frequency of DPWM and power stage.

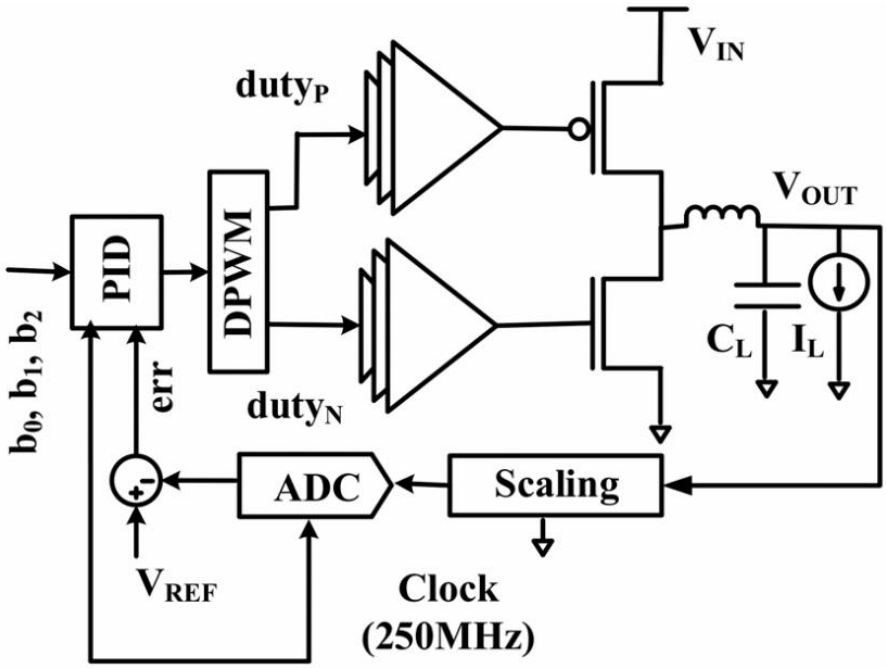


Figure 5.2: Architecture of IVR

The presented IVR design uses similar ADC design. Thus, the ADC transfer function will remain the same as Eq. 5.1. The compensator used is type III PID implemented in direct form. The z-domain transfer function for which is given by

$$H_C(z) = \left[\frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 - z^{-1}}\right] z^{-1} \tag{5.8}$$

The power stage is controlled through a zero-order-hold (ZOH). The continuous time transfer function power stage can be represented as

$$H_P(s) = G_{vd0} \frac{1 + \frac{s}{\omega_{ESR}}}{1 + \frac{s}{Q\omega_0} + \frac{s^2}{{\omega_0}^2}} \tag{5.9}$$

$$\omega_{ESR} = \frac{1}{ESR_C C} \tag{5.10}$$

$$\omega_0 = \frac{1}{\sqrt{LC}} \tag{5.11}$$

$$Q = \frac{1}{ESR_L + dR_{PFET} + (1-d)R_{NFET} + (ESR_C \| R_L)} \tag{5.12}$$

where d is duty cycle. Since the poles and zeroes are fixed for ADC and compensator similar to that discussed for DLDO system, the loop stability and response time for IVR will depend on power stage transfer function. From Eq. 5.9-5.12 we can observe that the small signal behavior of power stage is represented as a second order system with resonant frequency ω_0 and Q-factor (Q). Thus, we can observe that at ω_0 the loop frequency response is dominated by LC poles.

5.2 Analysis of NBTI Effect On On-Chip VRs

DLDOs and IVRs use PFETs for the power stage to perform DC-DC conversion of the voltage (Fig. 5.1, 5.2). NBTI can induce threshold voltage (V_{TH}) and mobility shift in PFETs. This shift in V_{TH} due to NBTI is considered as result of the generation of interface traps at the Si/SiO2 interface when there is a gate voltage. V_{TH} increases when electrical stress is applied and partially recovers when stress is removed. This process is commonly explained using a reaction-diffusion (R-D) model [48]. This results in an increase in 'ON' resistance of PFETs. Effect of such resistance shifts in power stage on the transient response time (settling time following a load step) for DLDOs and IVRs is anaalyzed in this section.

5.2.1 NBTI Simulation Method

The V_{TH} increase for the simulation analysis is modeled as increase in the on-resistance (R_{ON}) for PFETs in the power stage for both the designs as shown in Fig. 5.3. The DLDO and IVR designs are simulated using Simulink models based on equations from Section 5.1. The V_{TH} shift is estimated from predictive models [49] and shift in R_{ON} values are

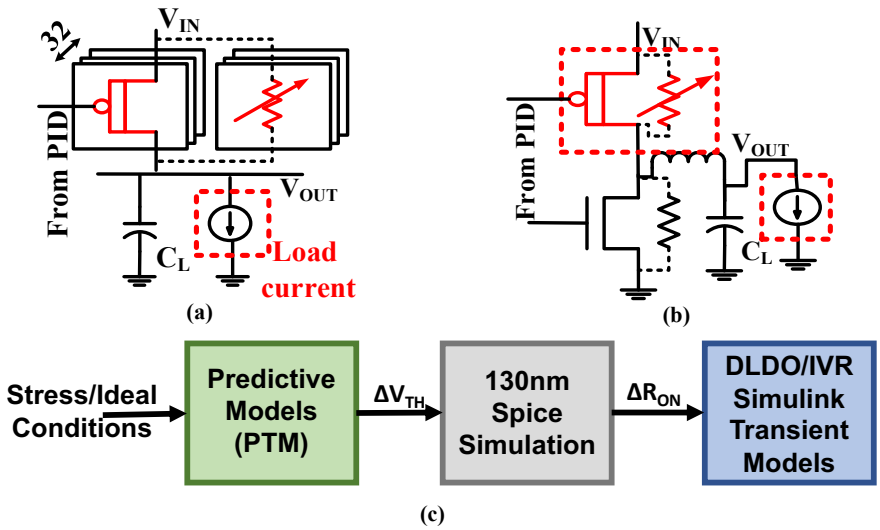


Figure 5.3: NBTI induced power stage aging simulation setup for (a) DLDO and (b) IVR power stage. (c) Simulation flow for power stage stressing

Table 5.1: V_{TH}, R_{ON} and F_{SAMP} shift using predictive models [49] and spice simulation for 130nm CMOS process

Voltage Stress* (V_{GS})	ΔV_{TH}	ΔR_{ON}	ΔF_{SAMP} (DLDO&)	ΔF_{SAMP} (IVR+)
2.0V	59.8mV	7.56%	-10.3%	-9.6%
2.4V	62.4mV	8.25%	-10.7%	-10.2%

* Stressed at room temperature for 10000s; The 130nm devices used are rated at 1.2V
& Controller critical path has 136 gates for DLDO system
\+ Controller critical path has 118 gates for IVR system

mapped to these from SPICE simulations for the devices in 130nm CMOS process. Table 5.1 shows the V_{TH} and R_{ON} shift estimated from the predictive model. The stress duration of 10000s and the stress levels (2V and 2.4V) are selected based on NBTI experiments performed in [50] for same 130nm process node.

Effect on DLDO

In response to a load transient, an LDO regulates output voltage by changing the resistance of the power stage. In DLDO, a power stage is designed by using an array of power PFETs, and regulation is achieved by controlling the number of 'ON' PMOS devices. The variation

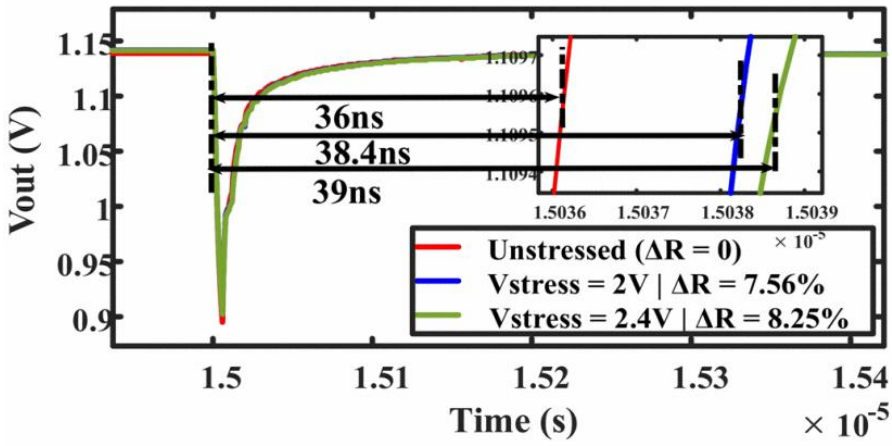


Figure 5.4: Simulated transient response for different stress levels for DLDO

in V_{TH} due to NBTI leads to increase in R_{ON} of individual PFET devices being stressed and decreases the current capacity of each PFET device (I_{PMOS}). This results in more PFET devices being ON before and after load jump to supply the same load currents. From Eq. 5.3-5.6 we can observe that open loop DC gain of the DLDO system is dependent on I_{PMOS} and will be reduced due NBTI stressing. This decrease in the DC gain of system will lead to reduced bandwidth which results in increase in response time. The controller can compensate for voltage error to ensure proper regulation but not for degradation in response time due loss in DC gain without dynamically updating the PID gains. From Eq. 5.3-5.5 we can observe that, the load pole for DLDO closed loop system is sensitive to R_{ON} of the PFETs; and a shift in the R_{ON} of the PFETs modulates the closed loop transfer function. Consequently, the simulation results show that DLDO suffers from a significant increase in settling time after a transient event when power stage is stressed. Fig. 5.4 shows a 6.7% and 8.3% increase in settling time for ΔR_{ON} of 7.56% and 8.25% respectively.

Effect on IVR

An IVR power stage has a pulse width modulated (PWM) signal as an input. The DC-DC conversion is achieved by duty-cycling the 'on-off' period of the power stages, and regulation is performed by changing the duty-cycle. The duty-cycle of the PFET device

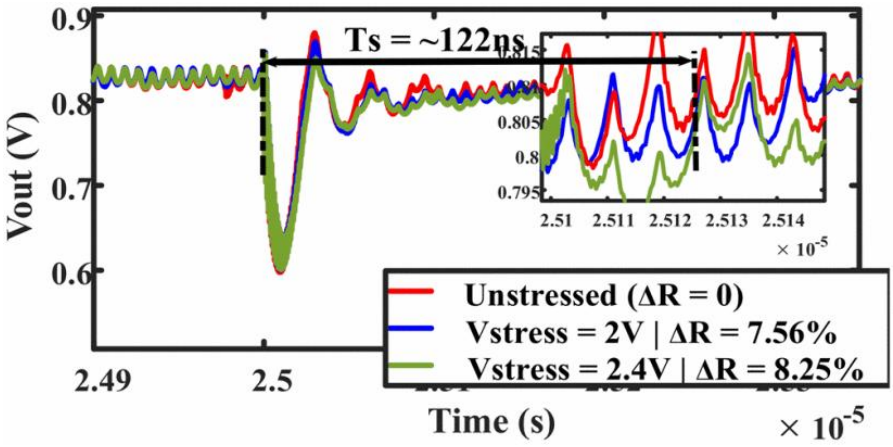


Figure 5.5: Simulated transient response for different stress levels for IVR

naturally reduces the effects of NBTI aging for same stress level. Moreover, from Eq. 5.9-5.12 the frequency response of the power stage is dominated by the LC poles which are relatively insensitive to R_{ON} of power stage PFETs. Consequently, a change in the power stage resistance has a very limited effect on transient response, and as observed in Fig. 5.5, simulations show minimal effect on the settling time due to load transient events.

5.3 Tuning Against Aging-Induced Degradations

Since on-chip voltage regulators are fabricated on the same die as digital core, they are expected to undergo similar aging degradations and process/passives variations. Thus, a tuning engine enabling post-silicon tuning against the aging and variations in process and passives will be very helpful in improving the transient response of a system after aging. An auto-tuning engine based on [18] has been implemented in the 65nm test chip to demonstrate the ability to improve transient performance of a DLDO based system post aging.

The tuning engine is implemented in time-domain for light weight design and low complexity. The digitized error signal err (Fig. 5.1) is used as a tuning metric and generates optimal compensator gains to minimize a cost function. The cost function is defined as a weighted summation of absolute error (AE), signed error (SE) and settling time (CT) as

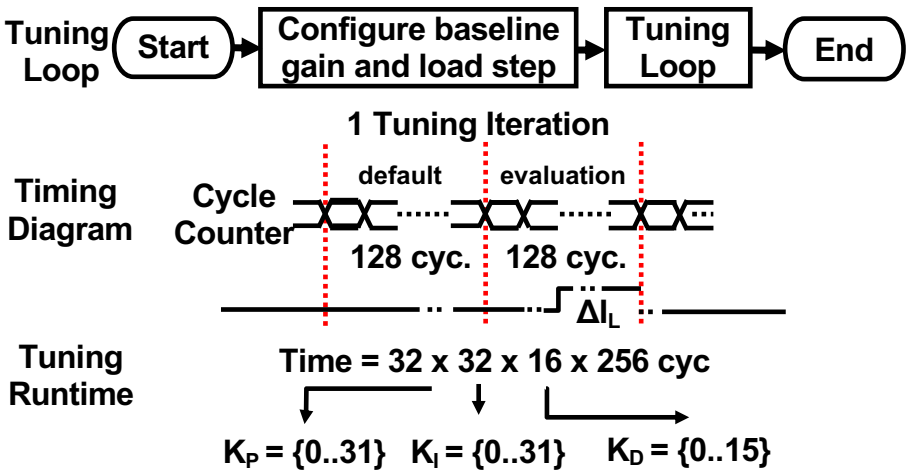
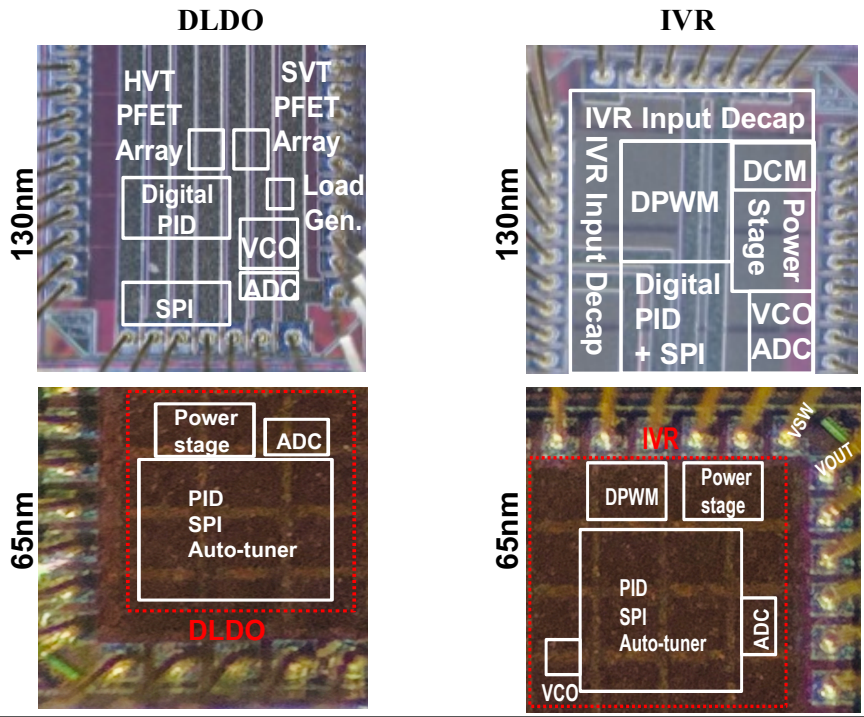


Figure 5.6: Control flow of the DLDO auto-tuning algorithm [18]

shown in Eq. 5.13.

$$cost = \alpha \times AE + \beta \times SE + \gamma \times CT \tag{5.13}$$

AE accumulates the absolute value of the digitized error signal and is used to eliminate unstable responses. SE accumulates signed values of the digitized error and is mainly used to capture damped responses providing higher phase margin. Finally, CT is defined as number of cycles it takes for the error to become less than a threshold, determining the settling time. Based on the application, different weights can be selected for the cost function leading to optimal system configuration. Fig. 5.6 shows the control flow for the tuning algorithm. The cost for each compensator gain configuration is computed in an evaluation period preceded by a default period. In the middle of the evaluation period, a load transient is induced via on-chip load generators to capture error patterns for low-high load transitions. The load resets to default value in default phase and a baseline gains are loaded. This ensures same initial conditions for all the gain configurations. A full sweep across all PID gains results in 16.7ms time for the tuning. It can be further reduced by using targeted upper and lower limits for the gain ranges. One-time post-silicon tuning is performed to mitigate process variation impacts, while infrequent online autotuning is performed to mitigate aging-induced degradations.



	Process	F_{SAMP} (MHz)	V_{IN} (V)	V_{OUT} (V)	L (nH	C (nF)
DLDO	130nm [18]	250	0.5-1.22	0.35-1.17	-	1.9
	65nm [54]	250	0.6-1.2	0.4-1.13	-	3.8
IVR	130nm [55]	250	1.25	0.5-1	11.6	3.2
	65nm [54]	250	0.9-1.2	0.6-1	62	23

Figure 5.7: Test chip micrographs and design specifications

5.4 Measurement Results

We verify the trends predicted from simulations using measurement of DLDO and IVR test-chips designed in 130nm and 65nm CMOS process. Due to the design limitations of the test chips, measurements are performed for only power stage aging. For the test chip measurements, the accelerated aging degradations in the power stage are induced by applying a voltage stress for 10,000 seconds [50] and performing measurements after reverting the supply to the nominal operating conditions. While stressing, all the PFETs are forced to switch ON by forcing the drain of PFETs with an external load current. Fig. 5.7 shows the test chip micrographs along with design details for both systems and Fig. 5.8 demonstrates the measurement setup for stressing the devices along with analysis of degradation

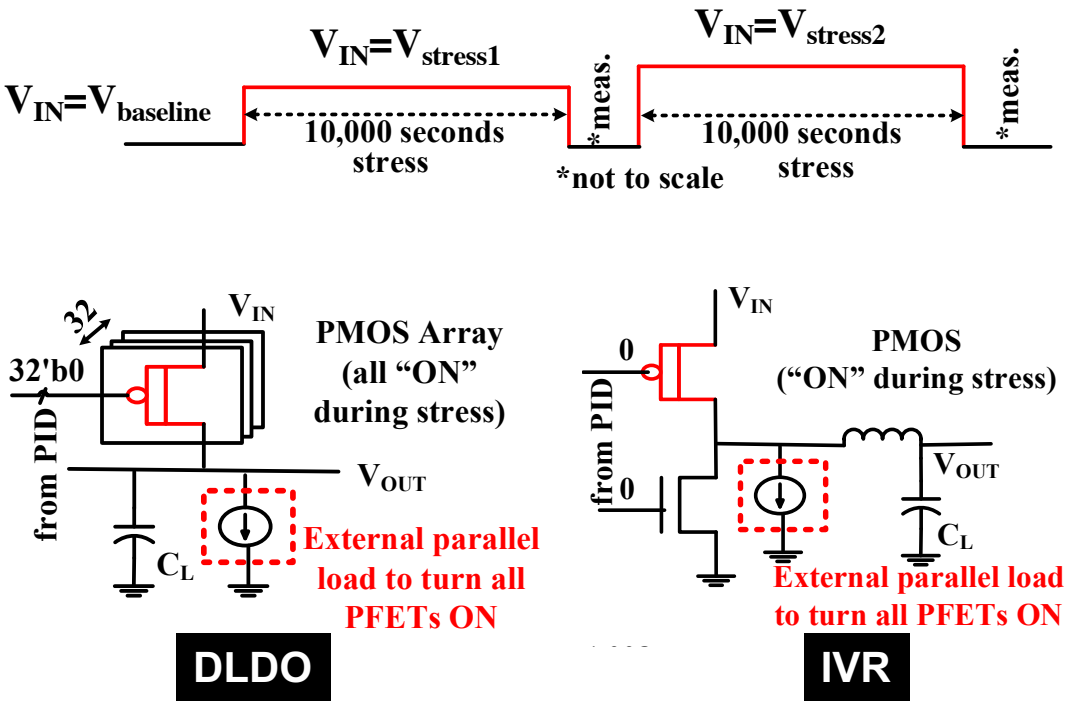


Figure 5.8: NBTI induced power stage aging measurement setup

in transient performance. Stress levels of 2V and 2.4V are implemented on the 1.2V rated PFETs in 130nm process and 1.5V to 1.9V are implemented for 1V rated devices in 65nm process.

5.4.1 Effect on DLDO

The measurement data follows similar trend as simulation results, but it is noted that there is a difference in performance degradation. This can be attributed to the fact that the predictive models do not include the effect of mobility change due to NBTI in the V_{TH} shift estimation. From Fig. 5.9a it can be observed that the voltage levels before the load step and the droop values change when stress is applied. This is attributed to the fact that when V_{TH} of PFETs increase, the current through each PFET decreases resulting in more PFET devices in the array turning on to supply same current. Since we are operating at low drop out (60mV) with different number of PFET devices for stressed and unstressed system before the load jump, this may result in output to settle at different level of the same ADC bin. This results

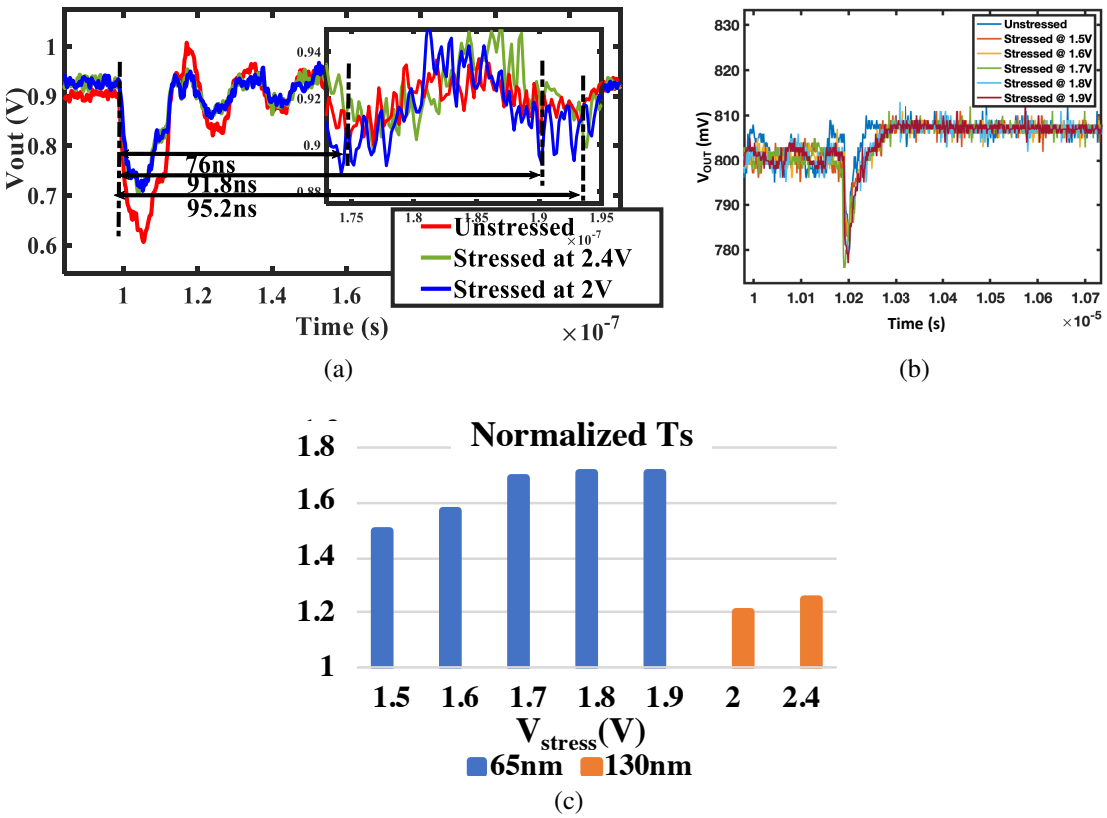


Figure 5.9: Measured transient response of DLDO under different stress levels for (a) 130nm process and (b) 65nm process. (c) Measured degradation in response time of DLDO system due to power stage aging

in different initial conditions for the stressed and unstressed system before the load jump and consequently effecting the droop values. From Fig. 5.9, a maximum degradation of 25.3% and 71.4% is observed in transient response time when the power stage is stressed at 2.4V and 1.9V for 130nm and 65nm test chips respectively due to the reduction in the bandwidth of the system due to stressing.

5.4.2 Effect on IVR

As predicted from simulation, the measurement results show that similar stressing on the IVR shows negligible change in the transient time. This can be observed in Fig. 5.5 and Fig. 5.10. However, IVR will incur a loss in power efficiency as it is dependent on R_{ON}. A maximum degradation of 0.65% and 3.2% is observed in power efficiency when the power

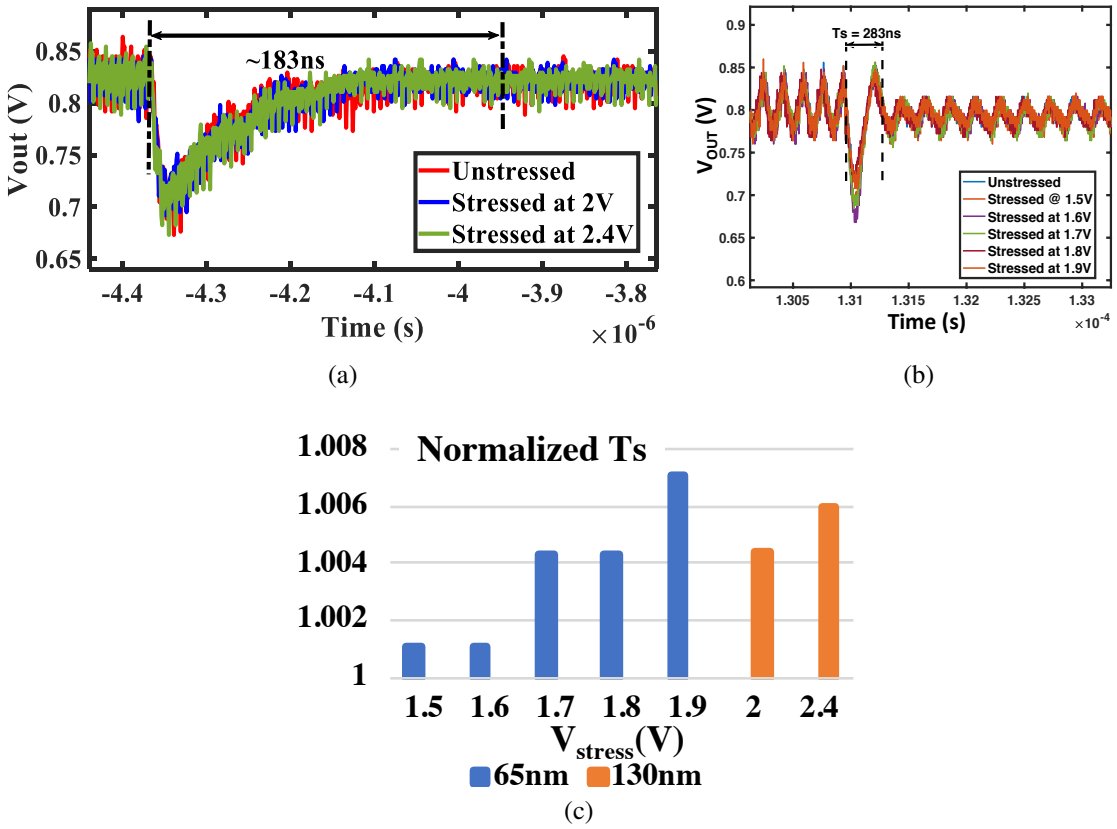
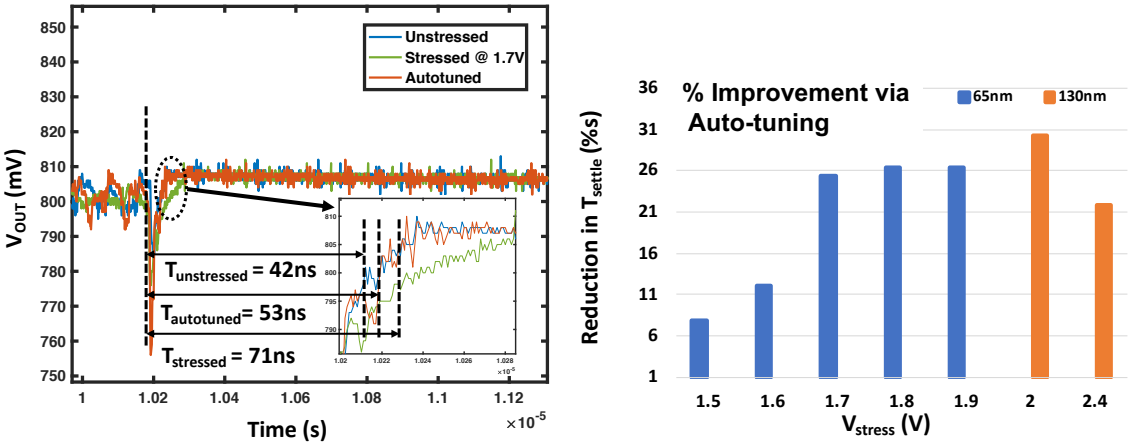


Figure 5.10: Measured transient response of IVR under different stress levels for (a) 130nm process and (b) 65nm process. (c) Measured degradation in response time of IVR system due to power stage aging

stage is stressed at 2.4V and 1.9V for 130nm and 65nm test chips respectively.

5.4.3 Tuning Against Aging-Induced Degradations

As expected, the accelerated stressing of the power stage leads to the increase in V_{TH} of the PMOS due to the NBTI effect. This results in change in closed-loop characteristics of the system (Section 5.1). Auto-tuning engine when enabled compensates for this shift in V_{TH} by retuning the compensator gains. Fig. 5.11a shows that in the 65nm test chip, on-line tuning of compensator gains reduce settling time by 25.4% compared to one-time postsilicon static tuning for DLDO system stressed at 1.7V. Fig. 5.11b illustrates detailed results for improvement in settling time for DLDO under various stress levels for both



(a) (b)

Figure 5.11: (a) Measured transient response in 65nm testchip for DLDO demonstrating 25.4% improvement in response time due to auto-tuning for aging induced degradations (b) Measured improvement via online auto-tuning in response time for DLDO system at various stress levels across 65nm and 130nm test chips

65nm and 130nm test chips. For 65nm and 130nm test-chip a maximum improvement of 26.1% and 30% is observed in settling time respectively due to auto-tuning. Since there is no significant change in response time due to stressing in an IVR, tuning is performed only for the DLDO.

5.5 Discussion

Since the controller is modelled based on state-space equations and not device equations, the V_{TH} increase due to NBTI cannot be modelled as increase in on-resistance. However, the V_{TH} increase can be mapped to increase in the critical path delay of the controller. Fig. 5.12a illustrates the simulation method used for stressing of the controller. The critical path for the controller is extracted from post place and route netlist. A spice simulation is performed on the netlist to measure critical path timing. The predictive models are used to estimate the V_{TH} increase due to aging and the spice models for the PFETs are modified to include the V_{TH} shift. The critical path is then retimed with updated spice models to observe the increase in the delay due to V_{TH} increase. The corresponding reduction in

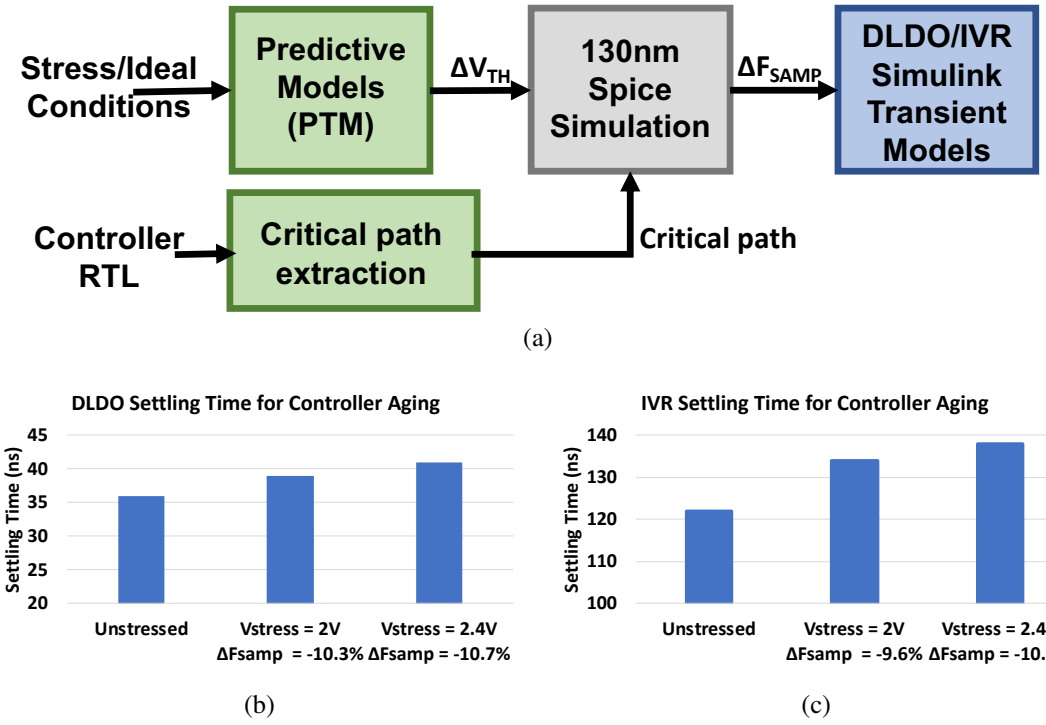


Figure 5.12: Simulation setup for controller aging in on-chip voltage regulators. Simulated transient performance degradation due to aging of feedback loop controller in (b) DLDO system and (c) IVR system

controller frequency is applied on Simulink models for transient analysis. Table 5.1 shows the V_{TH} and controller F_{SAMP} shift estimated from the predictive model.

The V_{TH} increase due to NBTI induced aging in the PFET devices in the controller will lead to increased critical path delay in the controller. To mitigate any timing errors in the controller, the feedback loop has to operate at a reduced frequency which results in ADC sampling error at lower rate. This will lead to slower reaction time towards any transient events thereby increasing the response time for both IVR and DLDO systems. Fig. 5.12b and Fig. 5.12c illustrate the increase in response time due to NBTI induced aging in controller for DLDO and IVR systems respectively. A maximum of 13.9% and 13.1% degradation in settling time is observed for DLDO and IVR based systems respectively.

In this work, the analysis, simulation and measurements performed for IVR and DLDO are mainly with a linear control algorithm. This was selected due to ease of design, qualitative analysis and integration applicable to linear control loops. It is well established that

for DLDO system transient response is highly dependent on control algorithm. Non-linear control generally provides better response but increases design complexity. Modelling the nonlinear control algorithms into transfer functions is not straight forward and will vary for different implementations. This leads to difficulty in estimating the effect of the V_{TH} and R_{ON} shifts qualitatively for overall closed loop DLDO system with non-linear control. Hence, we can't generalize the effect of NBTI induced aging DLDO power stage on transient performance as the effects might not be as pronounced with non-linear control.

5.6 Summary

Effects of NBTI induced accelerated aging of power stage and feedback loop controller of on-chip voltage regulators (DLDO and IVR) with linear PID control on transient performance (response time) and power efficiency are explored in this chapter. Simulations and qualitative analysis for NBTI induced aging of controller for both IVR and LDO indicate significant degradations in transient response time. In regard to NBTI induced aging of power stage, measurements from three test chips fabricated in 130nm and 65nm CMOS process demonstrate up to 25.3% and 71.4% degradation due to accelerated aging in response time following a load step for DLDO in respectively and almost negligible degradation for IVR. However, the IVR does incur some marginal degradation in power efficiency up to 0.65% and 3.2% in 130nm and 65nm test chips respectively. Thus, for on-chip voltage regulators with linear control, NBTI induced shifts in the power stage resistance has much smaller effect on IVR compared to DLDO. For DLDO systems with different nonlinear control loop, the effect of NBTI induced aging might be less prominent. Moreover a 26.1% and 30% improvement in response time against aging related degradations in DLDO power stage is achieved by auto-tuning in the 65nm and 130nm test chips respectively. Thus, online post-silicon tuning can be a key technique to improve reliability of DLDO against aging.

CHAPTER 6

AUTOMATIC GDSII GENERATOR FOR ON-CHIP VOLTAGE REGULATOR FOR EASY INTEGRATION IN DIGITAL SOCS

A modern processor/SoC requires multiple independent voltage domains to maximize energy efficiency through DVFS [7, 9]. Most often, a distributed power delivery architecture consisting of large global regulators like an IVRs powering smaller point-of-load (PoL) regulators like DLDOs are implemented to achieve those multiple power domains. However, the controller and power stages of on-chip voltage regulators for each voltage domain must be independently designed to match the target load demand i.e. maximum steady state power, power quality (voltage ripple) and transient (load/reference) performance. Since voltage regulators are typically a mixed-signal design, they usually need manual optimization and custom layout thereby increasing the design time and delaying time to market for SoC requiring on-chip voltage regulators.

This chapter presents an EDA tool (Fig. 6.1) for automated design and GDSII generation of two on-chip voltage regulators, mainly an IVR and a DLDO. And the integration of the generated on-chip VR within a digital SoC will significantly reduce the design time of VR-assisted SoCs. The key challenge for auto-generation of an VRs is to develop an EDA flow that couples the design of the control loop and the physical design of the controller and power stage to optimize the transient performance and efficiency of the VR. Such co-design is facilitated by integrating a front-end flow for frequency-domain design of the control loop (using MATLAB/SimuLink) to meet performance targets and a back-end flow for physical design of the power stage (using SKILL) to meet power demand. The integration is enabled by using an all-digital IVR and DLDO architectures that transform the designed control loop to a Register Transfer Language (RTL) realization and physical design to accurately consider the circuit level characteristics of the controller. The integrated

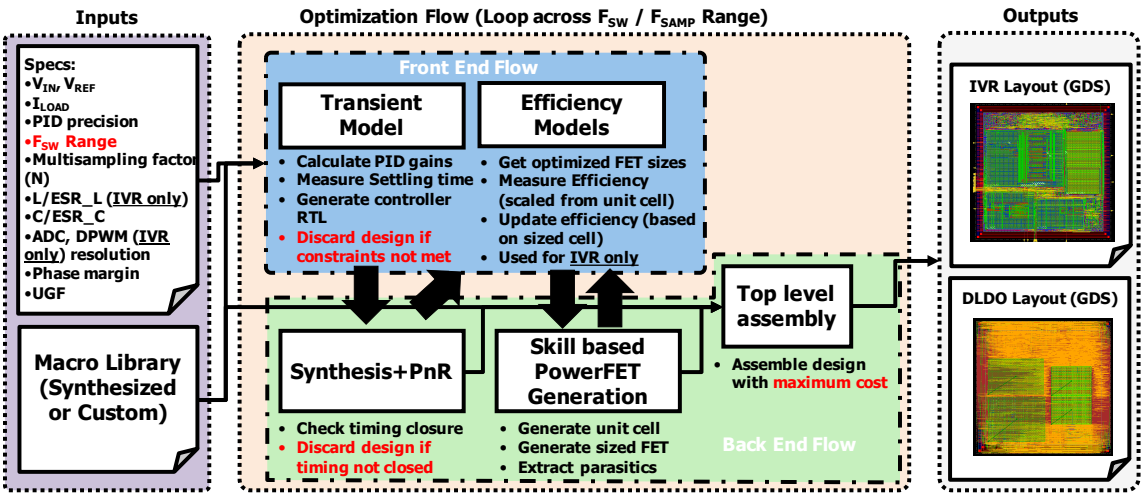


Figure 6.1: Specification to GDSII automation flow for an IVR and DLDO

flow is guided by an optimization method that selects appropriate configurations of the VR parameters to maximize transient performance and/or efficiency under constraints on power quality (voltage ripple). The output of the proposed flow is the layout and parameters of the IVR and/or DLDO for optimal performance/efficiency.

6.1 Overall Tool Flow

6.1.1 Front-end Flow: Behavioural Models

The front-end of the proposed flow is composed of (well-known) models of the IVR and DLDOs performance/stability (control loop) and efficiency as discussed below.

Baseline IVR Architecture

The IVR is implemented as an switched inductor voltage regulator. Fig. 6.2a shows an illustrative IVR architecture used to develop and demonstrate the proposed design flow. It consists of a power stage, that has PMOS and NMOS switches along with an LC output filter. The error at the output is sampled by the ADC at the rate $N \times F_{SW}$, where $N = 1$ in the case of single sampled systems and $N > 1$ for multi-sampled systems [1]. Multisampling

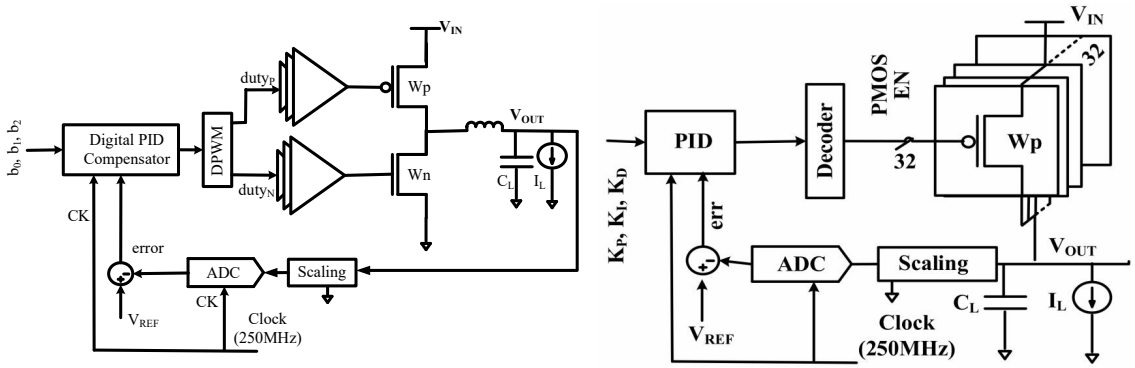


Figure 6.2: Simplified architecture of an (a) IVR and (b) DLDO

implements sampling frequency at a higher rate than switching frequency. This reduces the effective delay of the DPWM and improves bandwidth of the compensated system.

Baseline DLDO Architecture

The DLDO is implemented as an array of PMOS devices which are turned ON/OFF to perform regulation. Fig. 6.2b shows an illustrative DLDO architecture used to develop and demonstrate the proposed tool flow. It consists of a power stage, that has PMOS switches along with a load capacitor. Additionally, the voltage error is compensated using a PID controller.

Controller Model

The output of the IVR is sampled and digitized using an ADC. The digitized error is computed using the digital word corresponding to the target reference voltage. This error is passed on to the digital type III controller,which is of the form,

$$G_{comp}(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 - z^{-1}} \tag{6.1}$$

The output of the controller is the duty cycle command d[n], which acts as input for Digital Pulse Width Modulator (DPWM). The DPWM compares the duty cycle command

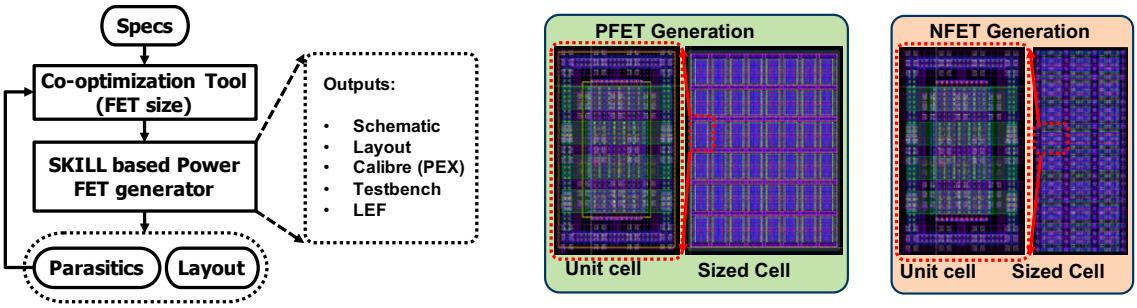


Figure 6.3: EDA flow for generating power stage

to a trailing edge sawtooth waveform with switching frequency F_{SW} to generate a PWM signal that drives the power stage, completing the loop.

Whereas, for the DLDO, the controller output goes to a decoder which determines number of ON PMOS devices and closing the control loop. The IVR power stage is modeled using state space equations to obtain the open loop transfer function [56]. And, the power stage of DLDO is modeled based on [26, 18].

Efficiency Model

In order to accurately calculate the efficiency of the systems, the individual sources of power consumption must be identified, and accurately modeled. For an IVR, the power is lost due to the inductor (P_L), capacitor (P_C), the power FETs (P_{FET}), and the PDN (P_{PDN}) are modeled according to the methods described in [57].

6.1.2 Back-end Flow: Physical Design

This section discusses the back-end flow for physical layout (GDSII) generation of the target IVR and DLDO design. This flow generates layouts for both digital controller block, analog power stage blocks, and the top level integration for the two and is mostly similar for both regulators.

Power FET Generation

The power FET generation flow is illustrated in Fig. 6.3. FET sizes are the only inputs needed for the cell generation and characterization. The generated cell includes schematic, DRC/LVS clean layout, post layout extracted (PEX) schematic, testbench to characterize the PEX schematic and LEF file for macro integration at top level. The layout generation follows a templated approach, where a fixed power stage template (non-cascoded) is used when designing unit cell. The unit cell is then instantiated into multiple rows and columns to meet the target load requirement. The generated layout also includes a built-in tapered power grid till specified metal layer (default is metal 4) and customizable aspect ratio. The flow can also be scaled across different process nodes with minimal change to underlying code as the minimum DRCs, layer and via information is obtained directly from the process techfile. The power stage drivers are synthesized using a design flow based on [58].

Controller Generation

The controller is implemented as a Finite Impulse Response (FIR) filter, and the physical design is generated using digital synthesis and place/route flow. The controller can be designed for different target bit precision and maximum frequency (= n$\times$switching frequency, where $n = 2$ for double sampled design in case of IVR) allowing trading off bit-precision of the coefficients (b_0, b_1, b_2), which controls the loop response (i.e. regulation and performance), with power stage switching frequency which determines efficiency. A case study regarding this trade-off is provided in Section 6.2.3.

Top level assembly

Once all the required modules are generated, top level assembly for the IVR or DLDO is performed using digital P&R tools. This involves generating a top level RTL for IVR or DLDO which contains digital and analog top level modules. RTLs for final controller design and power stage drivers are part of digital modules. Analog modules include auto-

generated power stage and macros such as VCO, ADC and DPWM which can be either selected from a custom/third party macro library or fully synthesizable macros based on prior works [59, 60, 61] due to modular structure of top level RTL. Floor planning stage can be done manually (optional) or by default use floor plan of legacy design as template. Routing blockages are placed over macros to ensure there won't be any metal shorts.

6.1.3 Integration of Front and Back-end Flows

The front and back-end design flows are integrated using an optimization flow that generates an IVR or DLDO designed to meet the target power and performance requirements defined by the user

Optimization goal

We propose a cost function that is a weighted summation of the normalized efficiency and performance of the IVR as shown below:

$$\max \quad Cost = \alpha E' - \beta T'_{settling} \tag{6.2}$$

$$\text{where } E' = \frac{E - E_{min}}{E_{max} - E_{min}}, T' = \frac{T_{settling} - T_{min}}{T_{max} - T_{min}} \tag{6.3}$$

$$\text{s.t.} \quad V_{ripple} < V_{ripple,max}, \quad \Phi > \Phi_{min} \tag{6.4}$$

$$F_c > F_{min}, T_{settling} \leq T_{max} \tag{6.5}$$

where, maximum voltage ripple ($V_{ripple,max}$), minimum phase margin (Φ_{min}), minimum crossover frequency (bandwidth) (F_{min}), and maximum settling time (T_{max}) are defined as optimization constraints. E_{max}, E_{min} and T_{min} are the maximum & minimum efficiency, and minimum settling time for the IVR, subject to the constraints. E' and $T'_{settling}$ are the normalized efficiency and settling time used to calculate the cost while α and β are the weights defined by the user.

For the DLDO system, the cost function only depends on normalized settling time (i.e.

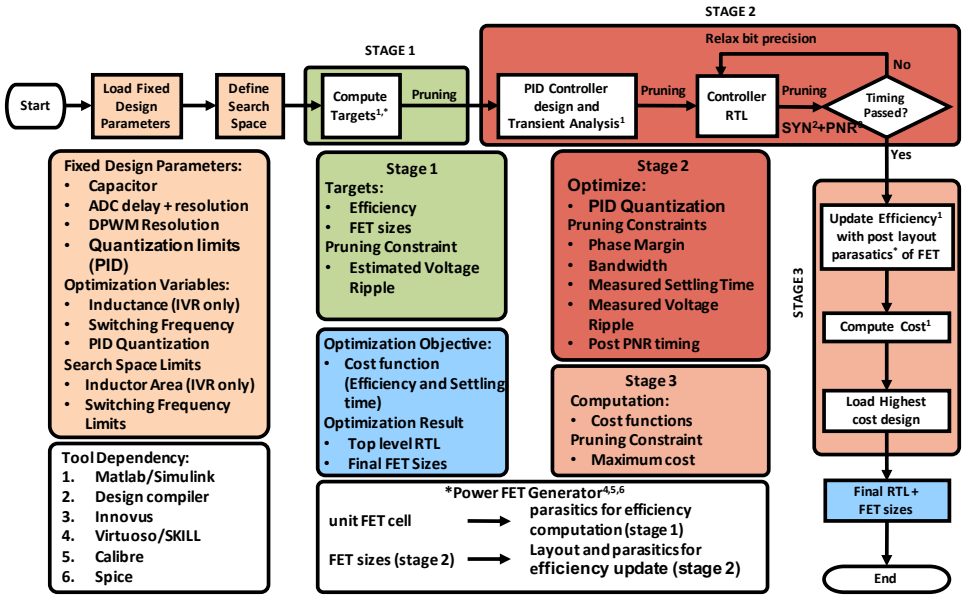


Figure 6.4: Efficiency and performance optimization flow

$\alpha = 0$ and $\beta = 1$) due to lack of current efficiecny models in front end. The future iterations of tool plans to include current efficiency models for DLDO in front end and account it in optimization goal.

Optimization method

Fig. 6.4 presents an overview of the flow used for controller design and co-optimization. To illustrate the flow, we assume a fixed capacitor size and inductance (IVR only) density. We consider inductance (IVR only), switching/sampling frequency and bit-precision of feedback loop coefficients as the control parameters.

The optimization flow follows a multi-stage design space pruning approach based on the defined constraints/targets. The initial search space is defined by switching/sampling frequency limits (derived from operating range of macros) and inductance (derived from area constraints for IVR design) set by the user. In the first stage of optimization, design targets including efficiency, power stage sizes and estimated voltage ripple for all the designs in the search space are computed using the front-end IVR models. While calculating

efficiency, parasitics for a unit cell of the power stage obtained from the back-end flow are scaled, and FETs are sized for maximum efficiency for the given load. The search space is pruned by filtering the designs using estimated voltage ripple as constraint. This is followed by Stage 2, where PID controller for designs in this reduced space are generated and transient analysis is performed on front end transient models to measure performance parameters such as measured voltage ripple and settling time. Search space pruning in stage 2 is determined by using phase margin, bandwidth, measured settling time and voltage ripple obtained from transient simulation as constraints. For these designs, maximum resolution of quantization for accurate realization of the controller is determined based on ADC and DPWM (IVR only) resolutions and corresponding controller RTLs are generated. Additionally a post PNR timing based design filtering is performed for PID controller where if the designs do not meet the timing for a frequency value in this reduced space, the bit precision/resolution is reduced for the PID coefficients while maintaining the same dynamic range. The designs where the controllers do not meet timing or if the PID coefficient bit precision falls below the lower limits are discarded. In the third stage, the efficiencies for all designs in the final search space are updated using the sized power stage parasitics obtained from the back-end flow for IVR designs. These updated efficiencies along with the settling time are then used to compute the cost for IVRs and only settling time is used to compute the cost for DLDOs. Finally, the design with maximum cost from the reduced search space is selected.

6.2 Experimental Demonstration

In this section, we analyze run-time of the tool (Section 6.2.1) and demonstrate the application of the tool to automatically generate GDS-II of IVRs and DLDOs. We consider following cases: (i) only *generate* layout of an IVR/DLDO with pre-defined parameters (Section 6.2.2), and (ii) *optimization* of IVR parameters for performance/efficiency (Section 6.2.3). We next show support for technology scaling and SoC integration (Section

Table 6.1: Runtime analysis of the proposed tool flow

Tool Stages	IVR (DLDO) Runtime[&]	
	Generation Mode	Optimization Mode
Optimization Flow[*]	∼3 mins	∼3.1 hrs (∼27 mins)
Power Stage Generation	<1s	
Top Level Assembly	∼30 mins	
Integration with SoC	∼25 mins	

[*] Optimization search space consists of 725 (L & F_{SW} sweep) designs for IVR and 25 (only F_{SAMP} sweep) designs for DLDO generation

[&] Runtime is measured on an 8-core Intel Core i7-7700 processor with 32GB RAM

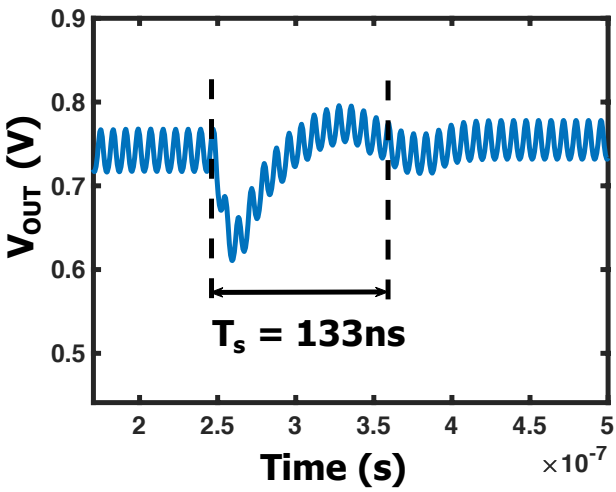
6.2.5) of the tool.

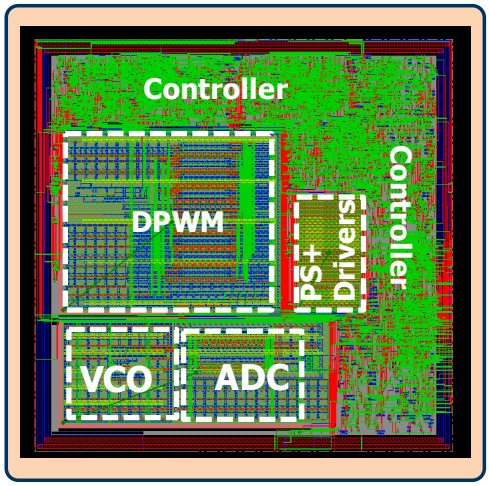
6.2.1 Runtime Analysis

Table 6.1 analyzes the runtime of different parts of the tool flow. The two operating modes (generation and optimization) of the proposed tool flow result in different runtimes. In the *generation* only mode, the top level P&R (back end) becomes the runtime bottleneck. For the *optimization* mode, the runtime will vary based on size and sweep variables of defined search space. A runtime analysis case considering coarse grain sweeps (ΔF_{SW}=4MHz, ΔL=2nH) leading to 725 and 25 possible designs for IVR and DLDO respectively is summarized in Table 6.1 . The optimization flow can be divided into three stages as explained in Section 6.1.3. It can be clearly identified that whenever optimization is performed the major bottleneck is the transient analysis and P&R performed in stage 2. Intelligently shrinking the search space can improve the runtime further.

6.2.2 IVR Generation for Pre-defined Parameters

In generation only mode, given all the design parameters such as conversion ratio, passives (LC), and switching frequency (F_{SW}), the proposed tool generates the IVR layout, and determines the controller coefficients to maximize performance/efficiency. We demonstrate this mode first by considering a custom IVR in 130nm CMOS [1]. To validate the





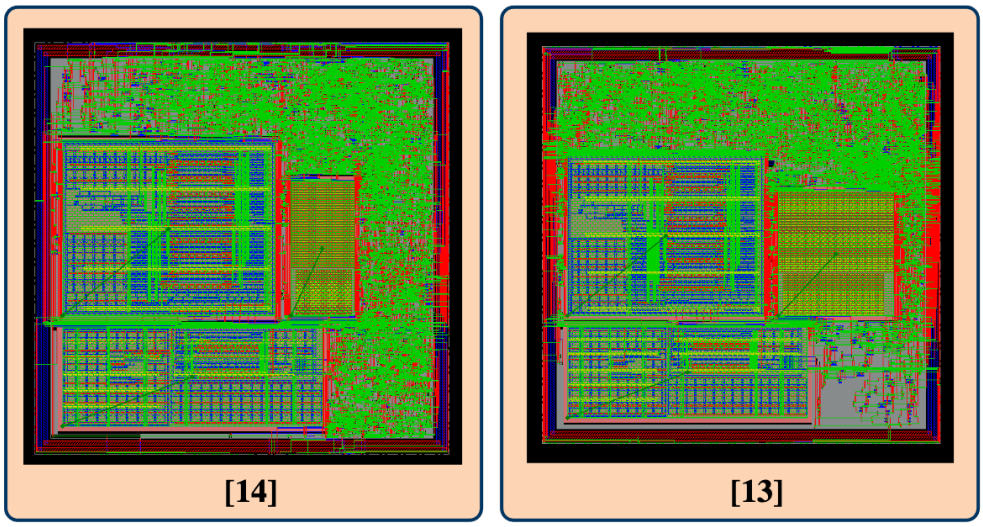
		[1]	This Work
Fixed Paramaters	**V_{IN} /V_{OUT} (V)**	1.2/0.75	
	L (nH) / C (nF)	11.8 (bondwire) / 3.2	
Optimized Parameters	**PID Quantization**	-	**6-bits**
Computed Targets	**Efficiency (%)**	71	**76**
	Settling time (ns) (ΔI_L =55mA)	200	**133**

Figure 6.5: IVR generated using proposed tool flow for specifications of [1]

design flow, we perform an analog-mixed-signal (AMS) simulation on the generated IVR including pad and package parasitics and compare it against the measured values reported in [1] as shown in Fig.6.5. It can be observed that the efficiency and settling time values of the generated IVR are comparable (marginally higher due to simulation and measurement mismatch) to those reported in [1]. We next apply the flow to generate IVR layouts in 130nm CMOS but based on specifications of designs in other technology nodes [14, 13] as demonstrated in Fig. 6.6. As expected, design [13] optimized for higher load current requires larger power stages.

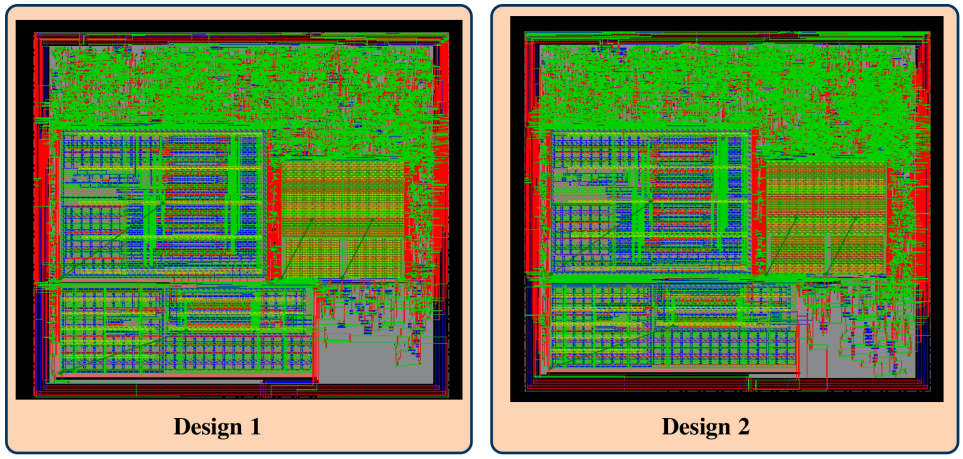
6.2.3 IVR Optimization: Case Studies

In this section, we discuss several case studies showing application of our tool to optimize different IVR parameters, and generate the final layouts.



	Fixed					Optimized	Computed		
	V_{IN}/V_{OUT} (V)	I_L (mA)	F_{SW} (MHz)	L (nH)	C (nF)	Quant. (bits)	Eff (%)	Ts (ns)	Area (mm^2)
[14]	1.5/1.0	180	500	1.5	10	6	80	12	0.176
[13]	2.0/1.2	800	500	3.08	1.83	6	74	157	0.158

Figure 6.6: IVRs generated in 130nm for specifications of [14, 13]



	Fixed					Optimized	Computed		
	V_{IN}/V_{OUT} (V)	I_L (mA)	L (nH)	C (nF)	Quant. (bits)	F_{SW} (MHz)	Eff (%)	Ts (ns)	Area (mm^2)
Design 1	3.6/1.0	1500	12	10	5	130	73.44	46	0.157
Design 2	3.6/1.0	1500	12	10	8	100	75.12	221	0.145

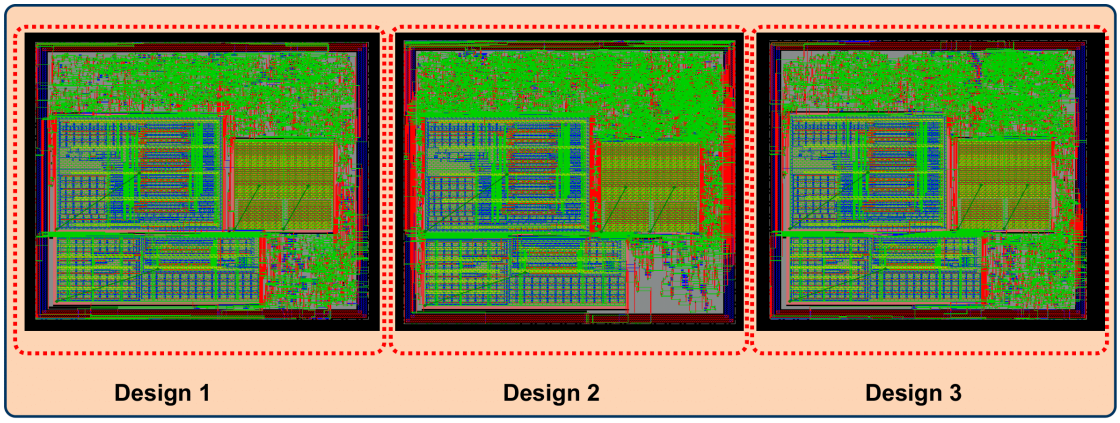
Figure 6.7: Quantization vs performance trade off

Controlling logic complexity for analog performance

A key aspect of our design tool is the ability to understand the trade-off between digital design complexity versus analog performance of the regulator. When implementing a digital controller the level of precision of the PID coefficients can determine the maximum operating frequency of the IVR. Higher bit precision for the PID coefficients results in a controller that is closer to an ideal continuous controller but will lead to a stricter timing budget. Reducing the precision can allow the system to operate at a higher switching frequency, which improves the settling time. Fig. 6.7 illustrates this trade-off between two designs using 5-bit and 8-bit quantization.

IVR generation for different optimization target

A key challenge in IVR design for SoC is to explore the trade-off space between performance and efficiency and develop design solutions to meet needs for different blocks. For



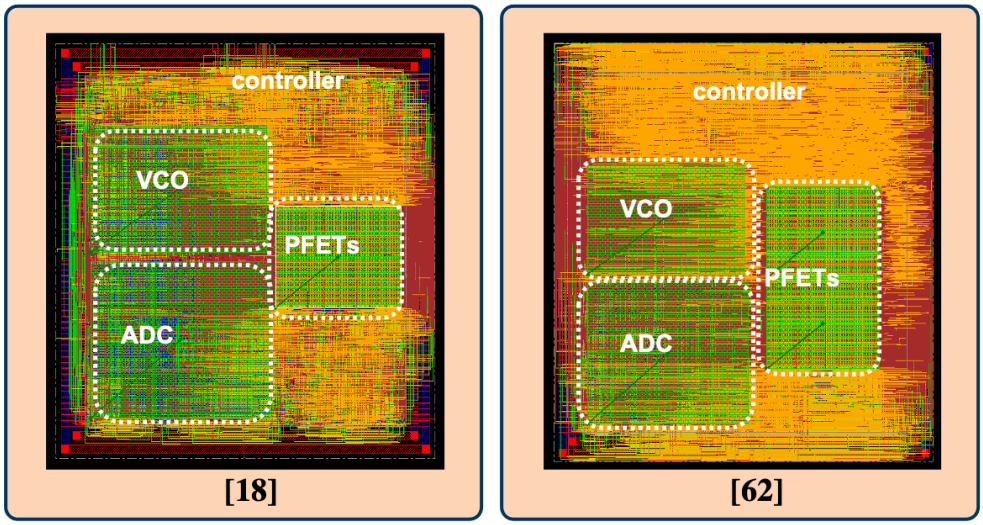
	Fixed				Optimized			Computed		
	V_{IN}/V_{OUT} (V)	I_L (mA)	C (nF)	Weights (α, β)	L (nH)	F_{SW} (MHz)	Quant. (bits)	Eff (%)	Ts (ns)	Area (mm^2)
Design 1 Eff. favored	3.6/1.0	1500	10	(1,0)	14	107	6	74.19	167	0.153
Design 2 Perf. favored	3.6/1.0	1500	10	(0,1)	27	125	6	69.42	14	0.161
Design 3 Balanced	3.6/1.0	1500	10	(0.7, 0.3)	16	117	6	73.05	27	0.158

Figure 6.8: IVRs with different optimization target

example, modules that generates large load steps or benefits from frequent DVFS transitions, prefer IVRs with faster transient response, while modules with steady power profiles prefer more efficient (but slower) IVRs. Our tool allows user to explore this trade-off space by selecting appropriate weights for the cost function defined in Eq. 6.2, and quickly generate corresponding layouts as illustrated in Fig. 6.8. It is observed that by selecting a design with small reduction in efficiency may lead to significant improvement in response time.

6.2.4 DLDO Generation for Pre-defined Parameters

Similar to generation only mode for IVR (Section 6.2.2), given all the design parameters such as conversion ratio, load capacitance, and sampling frequency (F_{SAMP}), the proposed tool generates the DLDO layout, and determines the controller coefficients quantization to maximize performance (transient response time). This mode is demonstrated by using the tool to generate DLDO layouts in 65nm CMOS but based on specifications of designs



	Fixed				Optimized	Computed	
	V_{IN}/V_{OUT} (V)	I_L (mA)	F_{SAMP} (MHz)	C (nF)	Quant. (bits)	Ts (ns)	Area (mm^2)
[18]	0.98/0.92	145	250	1.5	5	61	0.0569
[62]	1.1/1.0	210	250	20	5	300	0.0734

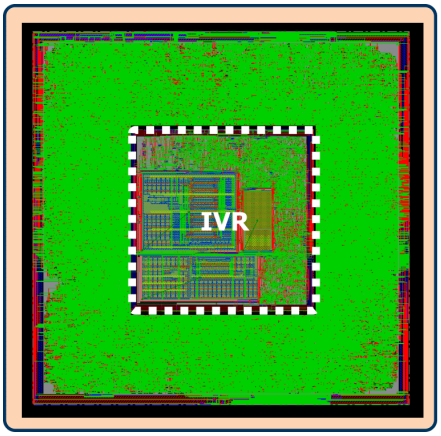
Figure 6.9: DLDOs generated in 65nm for specifications of [18, 62]

in other technology nodes [18, 62] as illustrated in Fig. 6.9. It can be observed that the tool is able to generate DLDO optimized for settling time, given target specifications with significant reduction in design time.

6.2.5 SoC Integration and Technology Scalability

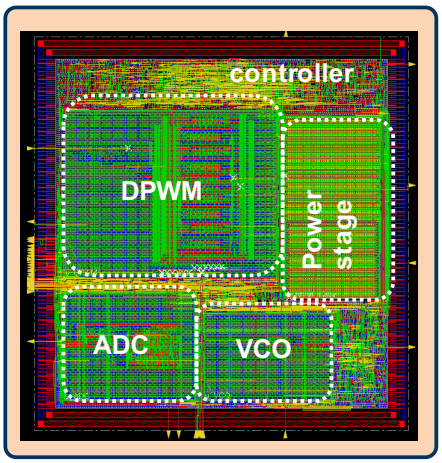
SoC Integration

The proposed flow facilitates an easy integration of the generated VR into a digital SoC. Consider a SoC with an IVR powering a RISC-V core. We generate a top level RTL of the SoC consisting of two modules, a core and the desired IVR. The RTL is then run through digital synthesis and P&R tools. During P&R it is ensured that the IVR module is defined as a partition and placed in center of the core for optimal power distribution. This is demonstrated by integrating a RISC-V core with an IVR in center for power delivery. Fig. 6.10 illustrates the layout and specifications for the generated IVR. We perform co-simulation of the core and IVR by generating a load profile for the cores using vector simulations using Synopsys PrimeTime. As expected, co-simulation shows that using an IVR results in improved transient response and voltage noise compared to an off-chip VRM when integrated with RISC-V core.



	Off-chip VR	**This Work**
Settling Time (ns)	2551	**109**
DVFS (V/μs)	0.29	**2.2**
Voltage Noise [V_{ref} - V_{min}] (mV)	80	**60**
Efficiency (%)	94	**81**

Figure 6.10: Integration of the designed IVR with RISC-V core



Process	65nm	130nm
V_{IN}/V_{OUT} (V)	1.2/1.0	
F_{SW} (MHz)	125	
L/C (nH/nF)	11.8/3.2	
Efficiency (%)	84.79	71
Settling Time (ns)	27	133
Area (mm^2)	0.0692	0.190

Figure 6.11: Scalability of the proposed EDA flow: IVR in 65nm

Scalability across technology nodes

The proposed flow supports scalabilty across various technology nodes. This can be attributed to the fact that front-end flow is independent of process node and the power stage generator flow (back-end) is scalable due to use of unit cell templates with minor changes in the generator code base. Controller and drivers are digitally implemented. The remaining macros can also be digitally implemented based on [59, 60, 61]. Thus, the tool can be migrated across process relatively easily. Fig. 6.11 illustrates an IVR designed and implemented in 65nm process.

6.3 Summary

This chapter demonstrates a scalable EDA tool flow for fast GDSII generation of digitally controlled high-bandwidth on-chip voltage regulators. The proposed flow optimizes the control loop and power stage of an IVR and DLDO to achieve desired transient performance and/or efficiency, and generate the physical design (GDSII) of the IVR and DLDO that can be easily integrated with the RTL of an digital SoC. The auto-generated VR shows comparable performance with custom design, while enabling orders of magnitude reductions in design time. Realization of intelligent design space exploration and efficient al-

gorithms in future can minimize optimization time and use of digital synthesizable macros can reduce design complexity and scalability. Moreover, the future iteration of the tool can include added support more complex VR designs with different converter types, topologies and control schemes such as multiple phases, non-linear control, and cascode power stages. The current open source public release of the tool includes only the front end flow [63]. The overall flow including the back-end physical design flow will be released in future.

CHAPTER 7

ALL-DIGITAL FULLY SYNTHESIZED ON-CHIP VRS WITH FLEXIBLE PRECISION ARCHITECTURE

The multiple case studies and analysis of the designs generated by the auto-generation tool flow presented in Chapter 6 shows that there is indeed a huge benefit of having an automated tool flow since it reduces the design time in orders of magnitude. This chapter explores a fully synthesized IVR and DLDO architectures implemented using an automated design and GDSII generation tool flow discussed in Chapter 6. Unlike in [4, 5, 64], the proposed architecture is completely synthesizable and scalable with very minor changes required in the underlying code base of any EDA tool flow. Moreover, the IVR design includes a flexible precision and variable frequency feedback loop architecture to enable improvement in transient response at different load ranges. Additionally, the proposed architecture also includes a lightweight auto-tuning engine to mitigate dynamic variations and aging impacts [65, 1, 53, 66, 18]. Specifically, this chapter discusses the following key contributions:

- A fully syntheziable digitally controlled IVR and DLDO architecture that can be easily synthesized using standard commercial place and route tools
- Synthesizable architectures for conventionally analog/mix signal modules such as analog-to-digital converters (ADC), voltage controlled oscillators (VCO) and digital pulse witdh modulators (DPWM) and a corresponding macro generator flow to seamlessly automate the design and layout of these modules.
- A flexible precision and variable frequency feedback loop architecture for IVR design that enables enhanced transient performance during low-precision/high-sampling mode and ability to trade-off switching losses with transient performance due to the variable frequency operation.

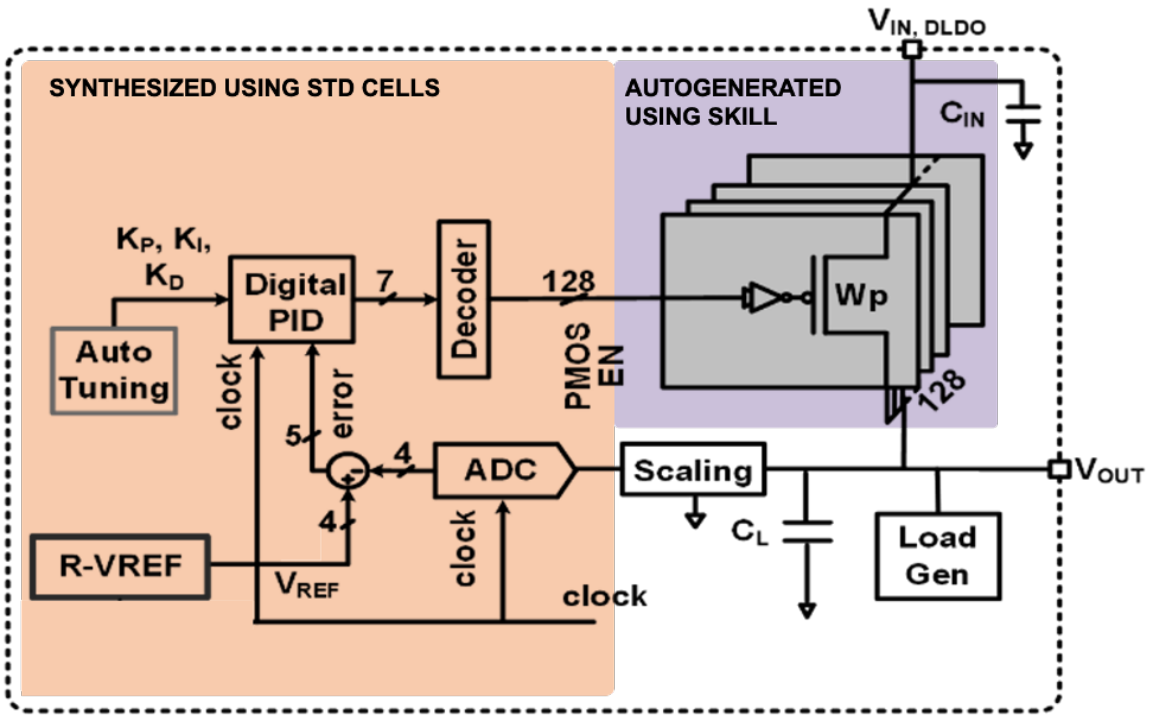


Figure 7.1: Overall architecture of synthesizable DLDO

7.1 System Design

In this section, we present the detailed system architecture of synthesizable DLDO, flexible precision IVR, and macro architecture.

7.1.1 Overall Architecture

Synthesizable DLDO Architecture

Fig. 7.1 illustrates the architecture of the proposed synthesizable DLDO. The DLDO power stage consists of 128 Power PFET devices in an array and a combination of on-chip MIM and MOS capacitance along with on board capacitance to form the output capacitance. A digitally controlled delay line based synthesizable ADC is used for digitizing the output voltage profile post scaling. The digitized output voltage is then compared with a digital reference word to generate a digitzed error. A digital PID compensator is used to compen-

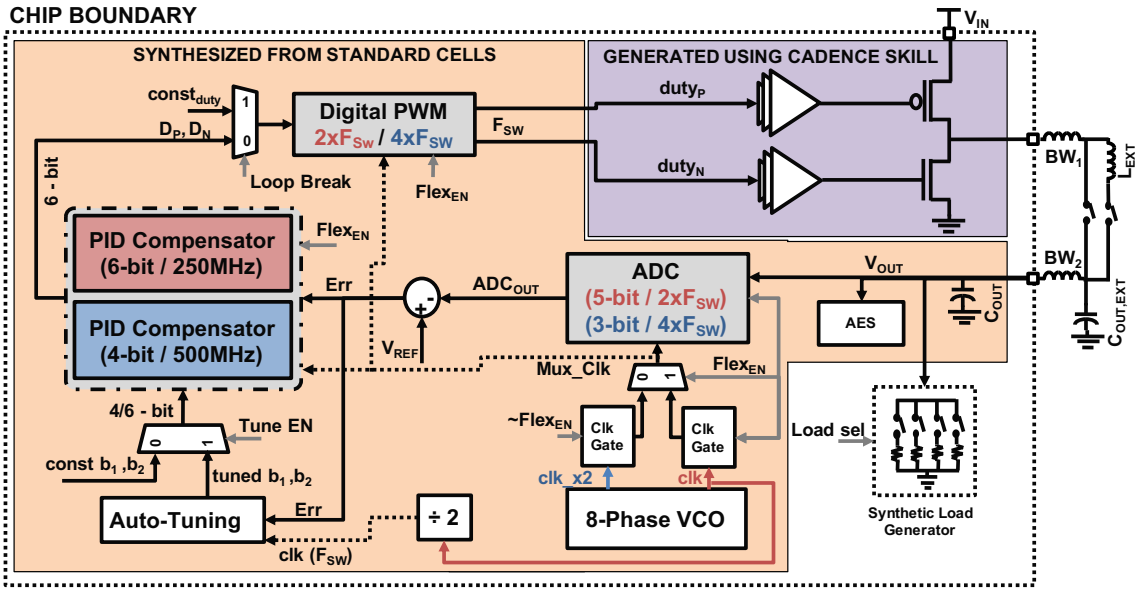


Figure 7.2: Overall architecture of the flexible precision synthesizable IVR

state for the voltage error and regulates the output by modulating the number of on PFET devices in the power stage.

Synthesizable IVR Architecture with Flexible Precision

The detailed architecture of the proposed synthesizable inductive IVR with flexible precision feedback loop is illustrated in Fig. 7.2. The IVR power stage output filter is implemented using combination of bondwire and diecrete inductances & on-chip and disctrete capacitances. The voltage error is captured similarly to the DLDO architecure using an ADC and a digital reference word. The IVR feedback loop is multi-sampled at a factor of switching frequency (F_{SW}) and includes two fully synthesized type-III (two zeros and one pole) proportional-integral-derivative (PID) compensator at different bit precision to compensate for the digitized error and generate a digital PWM control word. For both the digital PID compensator designs a type-III compensator implemented as following:

$$G_{COMP}(z) = \frac{\frac{b_0[n-1:0]}{16} + \frac{b_1[n-1:0]}{16}z^{-1} + \frac{b_2[n-1:0]}{16}z^{-2}}{1 - z^{-1}} \tag{7.1}$$

where b_0, b_1 and b_2 are each 6-bit and 4-bit digital words for high and low precision modes respectively assuming a fixed-point arithmetic calculation.

The digital pulse width modulation (DPWM) module then adjusts the duty cycle of the gate drive signals for power stage to regulate the output voltage, thus closing the control loop. A lightweight all-digital auto-tuning engine adapted from [3], is also implemented to perform post-silicon tuning of the direct form PID (b_0, b_1 and b_2) coefficients for both compensators to improve performance under passive and process variations.

The digital compensators, ADC, DPWM, the auto-tuning engine and a serial interface for programming are generated with digital synthesis tools. There are two closed loop paths in the proposed IVR architecture to facilitate the flexible precision operation. The selection of the closed loop path depends on an external configuration signal named $FLEX_{EN}$.

A digital voltage-controlled oscillator generates the two multi-sampling clocks for different precision modes which are then gated and muxed based on the $FLEX_{EN}$ configuration and then distributed throughout the control loop. The DPWM output clock (F_{SW}) is derived from the slower compensator clock to ensure that the duty cycle commands from the controller (D_N and D_P) change synchronously with F_{SW}.

7.1.2 Flexible Precision Operating Modes

The feedback loop is designed to work in two operating modes to enable the flexible precision operation. An external configuration signal $FLEX_{EN}$ is used to switch between the two operating modes. The first mode is defined as high-precision/low-sampling (HPLS) mode which is selected when $FLEX_{EN}$ signal is low. During this mode, the ADC and PID compensator ave high bit precision leading to better accuracy and sampling frequency as $2\times$ of power stage switching frequency (F_{SW}). This multi-sampling is enabled to improve bandwidth [1]. The other mode is selected when $FLEX_{EN}$ signal is high. This mode is described as low-precision/high-sampling (LPHS) mode where the ADC and PID now use a low bit-precision but instead sample at even higher rate of $4\times F_{SW}$. Our hypothesis is that

a higher multi-sampling operation with reduced precision of the coefficients (b_0, b_1, b_2) and ADC will help the IVR to meet a tighter performance constraint indicating a higher bandwidth of the loop. This will in turn lead to faster response to transient events.

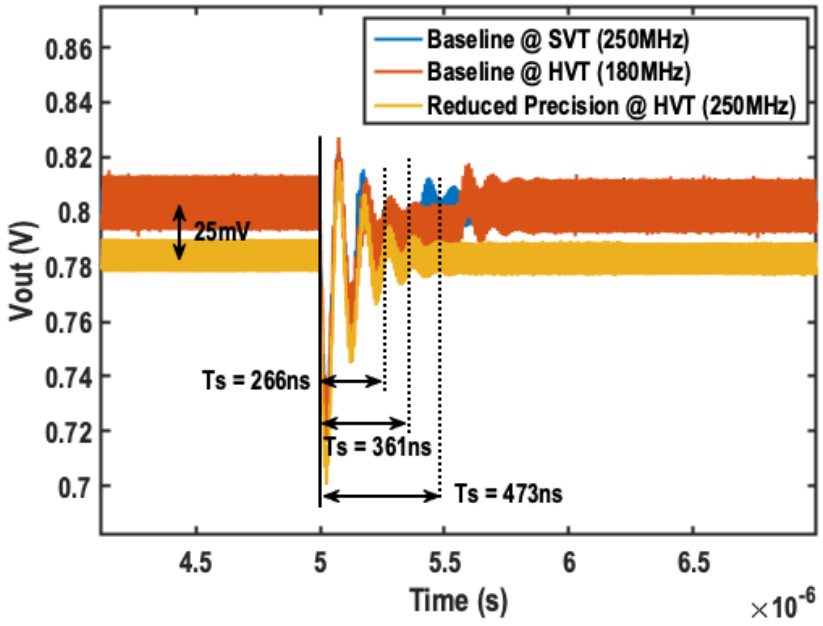
The default mode for feedback loop is high precision mode. The ADC, DPWM, and PID compensators are first synthesized for a higher bit-precision that can achieve timing closure for the target sampling frequency. The macro architectures are designed such that they are capable of running at higher sampling rate and lower precision during operation of the chip to achieve the LPHS mode. During both the modes, the DPWM converts control words generated at both $2\times F_{SW}$ and $4\times F_{SW}$ to a fixed F_{SW} frequency duty cycle signals for power stage. Moreover, both the fast ($4\times F_{SW}$) and slow ($2\times F_{SW}$) clocks are both derived from same VCO and gated complimentarily based on an $FLEX_{EN}$ signal. The gated clocks are then muxed and the muxed clock is used throughout the feedback loop to ensure synchronous operation.

The advantages of the flexible precision feedback loop architecture are highlighted through two practical applications simulated in Simulink as follows:

Tolerating Variations in Feedback Loop

A critical challenge in designing circuits in nanometer digital process node is to tolerate process variation that affects performance of the digital circuits [30, 31, 32, 33]. As the IVRs are designed in the same process nodes, they are also expected to suffer from same variations [1, 28, 34, 35]. This results in variations in transient (load and reference) performance, resulting in higher uncertainty in the performance of the digital cores.

The most generic way to tolerate such process variations in the feedback loop is to reduce the overall switching and sampling frequencies to ensure timing closure. This in-turn would increase the loop delay and ADC conversion delay [1] and effectively reduces the overall bandwidth of the system. The reduced bandwidth leads to poor transient response. To avoid degradation of the transient response, we propose to run the IVR loop at same



	Baseline (SVT)	Baseline (HVT)	Reduced Precision (HVT)
F_{SW}/F_{SAMP} (MHz)	125/250	90/180	125/250
ADC/DPWM/PID (bits)	5/6/6	5/6/6	3/6/4
$T_{settling}$ (ns)	361	473	266
Droop (mV)	84	84	85
Phase Margin (^{o})	37.3	35.5	33.4
Bandwidth (MHz)	33.2	27.8	35.8

Figure 7.3: Improvement in transient response using reduced precision mode while tolerating HVT shifts in feedback loop

switching and sampling frequency as baseline system but at reduced precision to ensure timing closure for the slower devices due to HVT shift process variations. This would result in a smaller ADC conversion delay due to lower resolution and leads to higher bandwidth. Thus, resulting in better transient response at the cost of ADC binning accuracy. Fig. 7.3 illustrates an example of such case where a system has shifted from SVT to HVT process corner. The frequency drop due to the HVT process shift is determined by extracting critical path of the feedback loop and ensuring timing closure at reduced frequencies. As expected lowering the frequency to tolerate the variations has increased the response time by 31%. But, when running the feedback loop at baseline frequency and reduced precision with 3-bit ADC & 4-bit PID precision we observe 43.7% reduction in response

time when compared to the reduced frequency operating mode. However, we observe an overall DC drop of 25mV at IVR output. This can be attributed to larger ADC bins due to lower precision leading to voltage to settle at lower levels of the ADC bin. Fig. 7.3 also highlights detailed transient and stability parameters for the three cases.

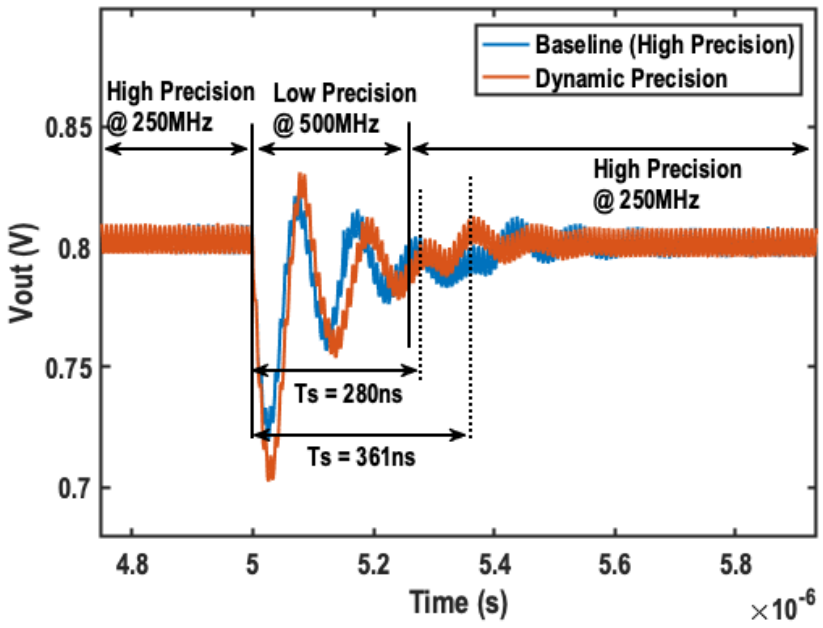
Non-Linear Control

Non-linear controllers are well known techniques to ensure fast transient response for IVRs. In the past there have been non-linear controllers for IVRs providing a resistive path to the output during transitions [1], or by non-linearly controlling gain of the feedback loop based on the voltage drop [18]. However, scalability of such non-linear controllers to a synthesizable architecture is difficult. Thus, we propose a non-linear control technique by modulating bit-precisions of feedback loop, and varying PID coefficients. The key concept is to use a higher precision during steady state operations while switching to a lower precision mode during transients based on user defined digital droop thresholds. The threshold values are chosen to ensure that the limit cycling does not trigger the dynamic precison control. Each precision mode uses optimally designed PID coefficients based on the auto-tuning discussed in Section 7.1.4. The bit-precision based, dynamic precision non-linear control is achieved using flexible precision designs of the feedback loop macros.

Fig. 7.4 shows that enabling the dynamic precision control helps reducing the transient response time for load step (20 to 120 mA) by 22.4% (81ns). However, the droop magnitude can be unpredictable due to larger ADC bins due to reduced precision as discussed previously. The frequency domain analysis ensures better bandwitdh and comparable phase margin to baseline high precision mode.

7.1.3 Synthesizable Flexible Precision Macro Designs

To enable a fully synthesizable architecture, the traditionally analog/mixed-signal macros in the feedback loop such as ADC, VCO and DPWM should have an architecture that can



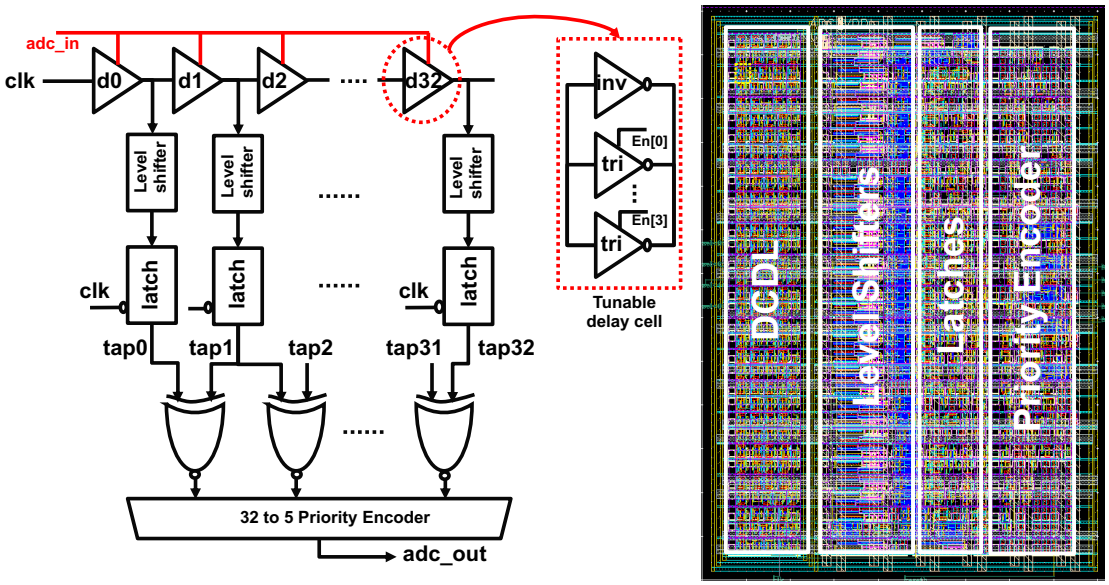
	Baseline (High Precision)	Dynamic Precision
F_{SW}/F_{SAMP} (MHz)	125/250	125/250 (125/500)
ADC/DPWM/PID (bits)	5/6/6	5/6/6 (3/6/4)
$T_{settling}$ (ns)	361	280
Droop (mV)	84	98
Phase Margin (^{o})	37.3	37.3 (32.8)
Bandwidth (MHz)	33.2	33.2 (34.5)

Figure 7.4: Improvement in transient response using dynamic precision non-linear control

be synthesized using digital place and route tools. Thus, to facilitate the synthesizablity and flexible precision the macro architectures are primarily based on digitally controlled delay lines (DCDL).

Analog to Digital Converter (ADC)

The ADC design illustrated in Fig. 7.5a is adapted from [67] and uses supply voltage of DCDL for the input voltage sensing (ADC_{IN}). The input sense voltage of the ADC controls the delay of each delay element in the DCDL since it is the supply for the delay line. Each stage in the DCDL consists of parallel tri-state inverters to allow post-silicon tunability of

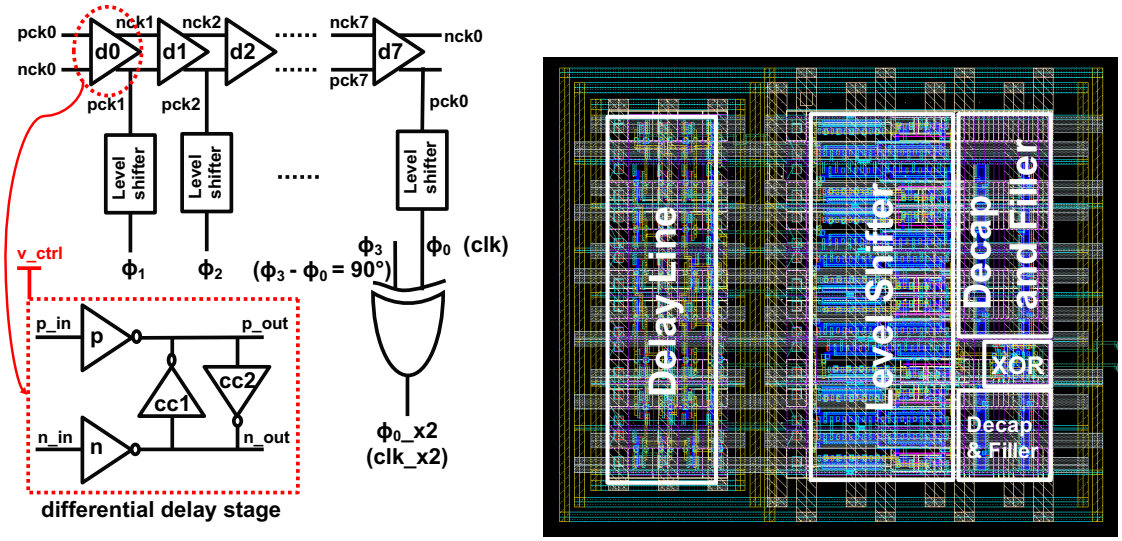


(a) (b)

Figure 7.5: (a) Detailed architecture of flexible precision synthesizable analog-to-digital converter (ADC) (b) Analog aware synthesized layout of the proposed ADC design

the delay line. The delay of each element can be adjusted by turning on/off parallel tri-state inverters using external configurations. The conversion cycle begins by sending clock signal at the input of the DCDL and ends when clock goes low. During the conversion time, depending on the delay of the individual elements, the input pulse crosses a partial number of delay elements, before clock goes low. Each delay element output is then level shifted and latched at the negative clock edge to sample and store the intermediate node when clock goes low. The level shifting is required since the intermediate node won't have full VDD swing and would be at ADC_{IN} voltage level. The latched outputs are finally converted to a 5-bit binary output through an XNOR logic followed by a 32-to-5 priority encoder. Since all the cells used in this architecture are available in foundry provided standard cell library, the design can be synthesized using digital synthesis tools. Fig. 7.5b shows the layout of the proposed ADC.

For the flexible precision operation, the ADC operates at 5-bit precision & samples at $2xF_{SW}$ in HPLS mode whereas it operates at 3-bit precision & $4xF_{SW}$ sampling rate at LPHS mode. To operate at LPHS mode, the clock signal starting the conversion cycle



(a) (b)

Figure 7.6: (a) Detailed architecture of synthesizable voltage controlled oscillator (VCO) with frequency doubler (b) Analog aware synthesized layout of the proposed VCO design

is now at $4xF_{SW}$ and thus will travel through only half the DCDL as compared to HPLS mode and thereby reduces the output precision from 5-bit (32 stages) to 4-bit (16-stages). Another LSB is shed from the output resulting in a 3-bit precision mode but maintaining same dynamic range.

Voltage Controlled Oscillator (VCO)

A fully digital 8-phase differential VCO is shown in Fig. 7.6a. The differential delay element is implemented using 4 inverters, where 2 inverters (p,n) are used to generate the complementary phases. The other 2 inverters (cc1, cc2) force the outputs of p,n to stay complimentary. The frequency tuning knob for the design is the supply voltage of the differential delay elements as it controls the delay of each element. The outputs of each of the delay elements are then level shifted to get the full VDD swing for the clock signal. This base clock is used for default HPLS mode and the power stage switching frequency is derived from this source by diving it by 2. For the LPHS operating mode, we require another clock signal synchronous with HPLS clock but at double the frequency. This double frequency LPHS clock is derived by performing XOR operation between the

base HPLS clock and a 90 degrees out of phase clock. The 90 degree out of phase clock is used for the XOR operation to ensure we obtain same duty cycle as base HPLS clock. Fig. 7.6b shows the synthesized layout for the proposed VCO design with minimal (1 XOR gate) area and power overhead.

Digital Pulse Width Modulator (DPWM)

The fixed-precision hybrid DPWM, motivated by [61] is designed to operate at flexible frequency to convert input of $2 * F_{\text{SW}}$ or $4 * F_{\text{SW}}$ to a fixed F_{SW}. The proposed design illustrated in Fig. 7.7 uses 2-bit counters and 32 stage DCDL to implement a 6-bit DPWM. Each delay element in the DCDL includes a parallel tristate buffers that are controlled by a digital delay locked loop (DLL) controller to control the delay of each element. For a traditional hybrid DPWM architecture the there are following constraints:

$$n = \log_2(\frac{F_{SAMP}}{F_{SW}}) \tag{7.2}$$

$$l = Int[\log_2(\frac{T_{SAMP}}{\Delta T})] \tag{7.3}$$

Where n is size of the counter and also MSBs of the control word used to compare the counter output to convert control word obtained at F_{SAMP} rate to duty signal at F_{SW} rate. And 2^l corresponds to the length of DCDL and ΔT is the delay of the individual delay element of DCDL. Thus the total DPWM control word width is defined as

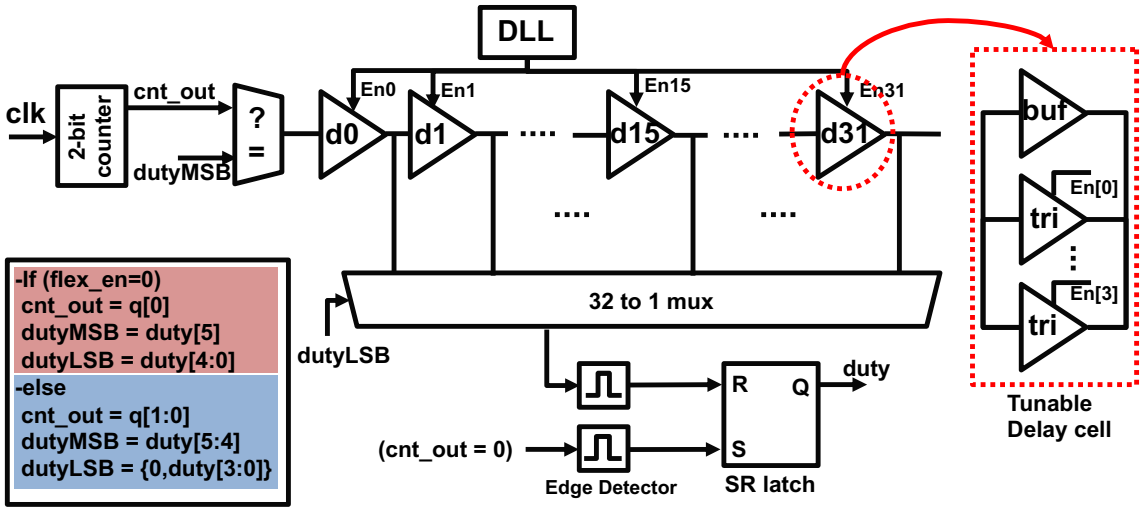
$$N = n + l \tag{7.4}$$

Where n corresponds to the MSB bits and l corresponds to LSB bits of the control word. Thus, for flexible frequency operation, the DPWM control word split should be

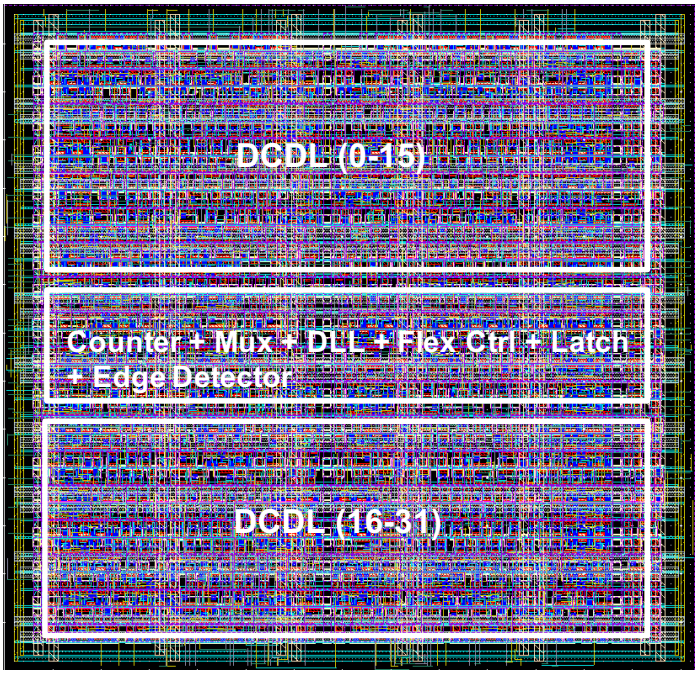
- 1-bit MSB and 5-bit LSB ($n = 1, l = 5$) for HPLS mode where $F_{SAMP} = 2 * F_{SW}$.

Thus, DPWM would require 1-bit counter and 32 stage DCDL for HPLS operation.

- Whereas, 2-bit MSB and 4-bit LSB ($n = 2, l = 4$) for LPHS mode where $F_{SAMP} = 4 * F_{SW}$. Thus, DPWM would require 2-bit counter and 16 stage DCDL for LPHS operation.



(a)



(b)

Figure 7.7: (a) Detailed architecture of synthesizable flexible frequency digital PWM (b) Analog aware synthesized layout of the proposed DPWM design

To minimize area and power overhead by having two separate DPWM blocks for both operating modes, the proposed design merges both the designs to have a single DPWM block with 2-bit counter and 32 stage DCDL and uses FLEX_{EN} control signal two switch between both the modes.

At HPLS mode when FLEX_{EN} is low, DPWM uses one MSB from the control word to compare the LSB of counter output $cnt\_out[0]$ to generate the pulse of divide by 2 CLK that propagates into the 32-stage DCDL. The 5 LSBs of the control word are used as select lines for selecting a delayed pulse from the 32 stages of DCDL. The DPWM output latches HIGH value when the counter resets to 0 and latches a LOW value when the selected delayed pulse goes high. During LPHS mode when FLEX_{EN} is high, 2 MSBs of the control word are used to compare the 2-bit counter output to generate the pulse of divide by 4 CLK that propagates through the DCDL. Compared to the HPLS case, since the input clock frequency doubled the pulse in the DCDL can only travel through half the stages. Thus, the LSBs of control are zero padded on the left to ensure that the delayed pulse is selected from the first 16 stages of the DCDL. The duty signal is generated by the same SR latch operation as done in the HPLS mode. The DLL controller locks the DCDL at either the full or half-length based on the FLEX_{EN} signal.

7.1.4 On-chip Auto-tuning

For an IVR control system, the use of integral of time-multiplied absolute error (AE) has been established to lead to an optimal transient response [1, 55]. The on-chip auto-tuning adapted from [1] consists of lightweight time-domain implementation (Fig. 7.8). The tuning algorithm uses a cost metric which is summation of aggregated absolute values of the digitized error values in the feedback loop. Fig. 7.8 illustrates the control flow of the tuning algorithm. During the tuning process, the cost is computed over 512 cycles consisting of a load and reference transient for different PID coefficient pairs in the search space and the pair with the minimum cost (aggregated absolute error) is selected. The PID

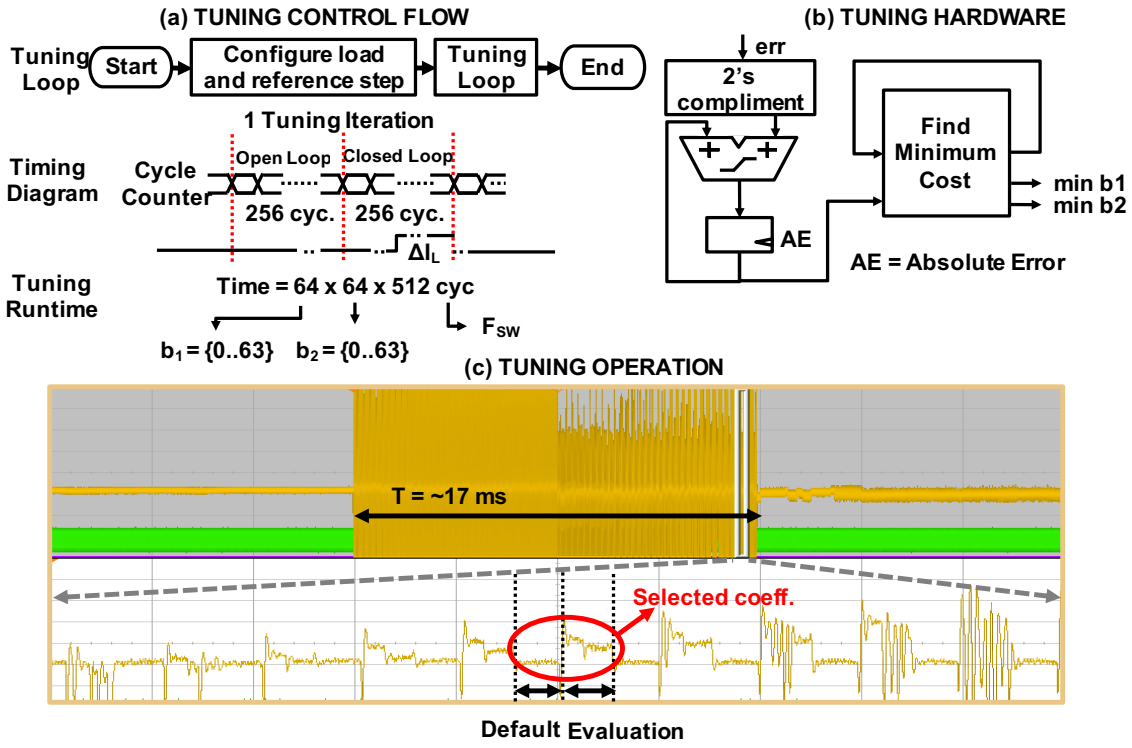


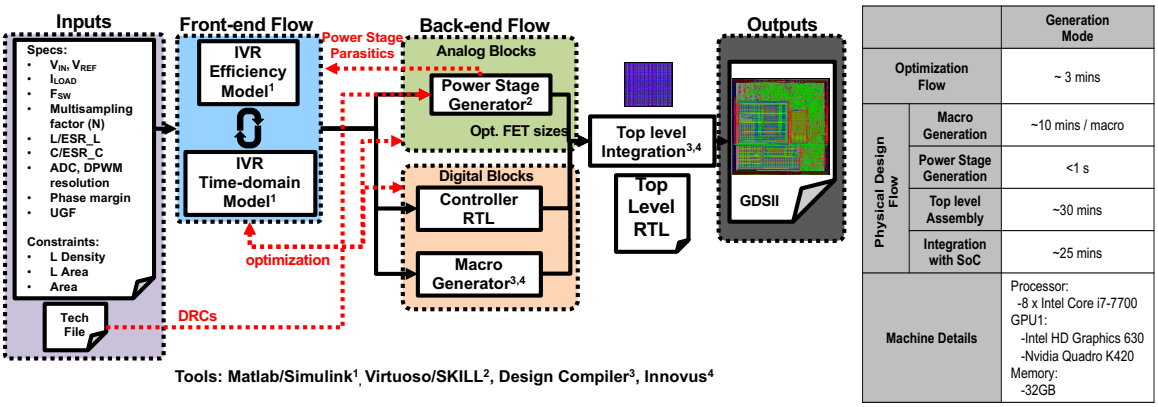
Figure 7.8: (a) Control flow of the on-chip auto-tuning engine (b) Hardware implementation of the tuning engine (c) Measured transient waveform of the tuning operation

coefficient pairs for both the compensators are obtained via the implemented tuning flow. A similar auto-tuning engine is also enabled in DLDO since the digital compensator is identical to that of IVR.

7.2 Auto-generation Tool Flow

7.2.1 Overall Flow

The GDSII of the IVR is generated using an automated tool flow adapted from [64]. The automated tool flow compliments the proposed synthesizable design by enabling rapid design and optimization process. Fig. 7.9 illustrates the overall automated tool flow. Given a target maximum output power, the power stage layout is auto-generated using multiple instances of a custom PCELL designed using Cadence SKILL by reading layout contraints directly from PDK to minimize process dependency. The fully synthesizable feedback loop



		Generation Mode
Optimization Flow		~ 3 mins
Physical Design Flow	Macro Generation	~10 mins / macro
	Power Stage Generation	<1 s
	Top level Assembly	~30 mins
	Integration with SoC	~25 mins
Machine Details		Processor: -8 x Intel Core i7-7700 GPU1: -Intel HD Graphics 630 -Nvidia Quadro K420 Memory: -32GB

Figure 7.9: Synthesized IVR/DLDO Auto-generation Tool Flow

including the compensator, ADC, DPWM, and VCO is implemented in standard cell based flow. The layout of the entire IVR is performed by automated place and route of the synthesized control circuits and auto-generated power-stage. The back-end flow is coupled with a front-end models of control loop (time/frequency - domain) and power stage efficiency using an optimization flow to determine PID coefficients for target bandwidth and phase margin constraints considering layout effects like power stage parasitics and maximum frequency of compensator.

7.2.2 Macro Generation Flow

To enable a fully synthesized system architecture, the feedback loop macros are implemented as synthesizable digitally controlled delay line (DCDL) based designs. To simplify and automate the process of auto-generating the macros and integration in the overall flow, a macro generation flow as illustrated in Fig. 7.10 is implemented. The first stage of the macro generation flow takes input parameters such as macro type, architecture, resolution and list of standard cells to be used. Once the input parameters are specified, a gate level RTL is generated using RTL templates based on the specific macro and architecture selected. A power intent file is also generated based on the same templates to ensure proper power planning for the multiple supply voltage (MSV) architecture of the selected macro. The gate-level RTL and power intent file are then run through standard synthesis and place

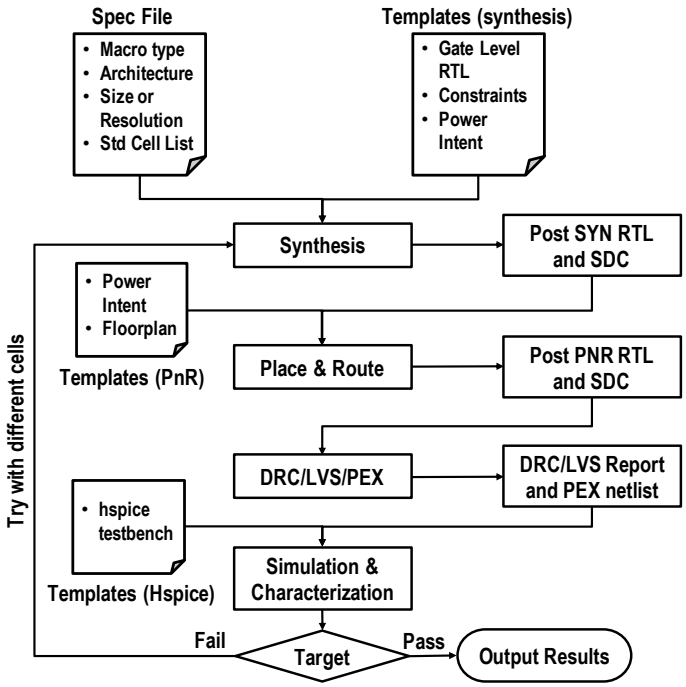


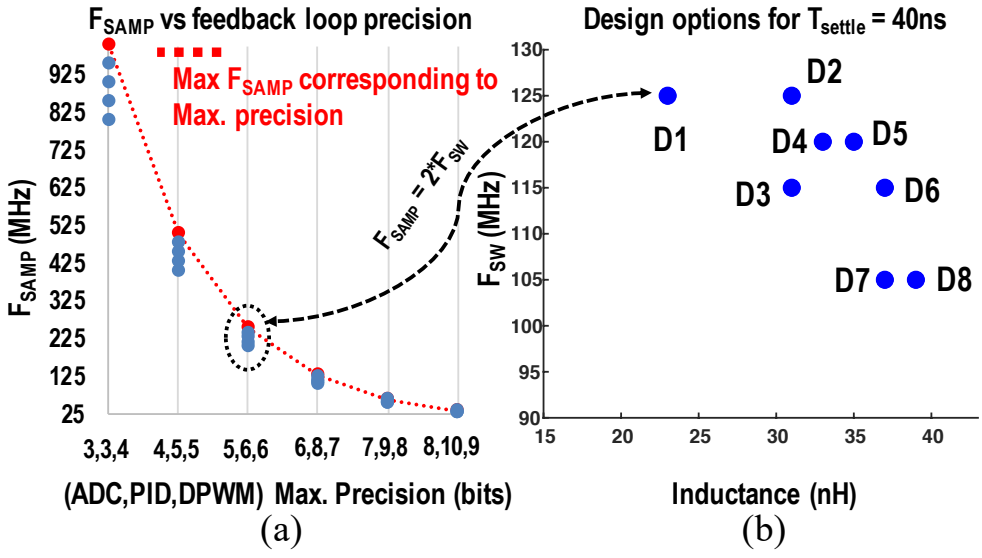
Figure 7.10: Macro generation tool flow implemented in the automated IVR/DLDO generation tool flow

and route tools to generate a layout and gate level netlist for the macro. During the place and route stage it is ensured that the delay cells are placed in separate isolated power domain and floorplanned to have symmetric and balanced arrangement (Fig. 7.5b, Fig. 7.6b, Fig. 7.7b) based on legacy template layouts. The generated layout and netlist is then imported to Cadence Virtuoso to verify DRC and LVS along with parasitic extraction (PEX). The PEX netlist is then used to verify and characterize the macro. If the targets are not met, then the process repeats with different sized standard cells from the list until the targets are met post PEX. Currently the flow supports only VCO, ADC and DPWM modules using a DCDL architechture as discussed in Section 7.1.3. But due to the simplicity of the flow, expanding the capability of the flow to support multiple macro deisgns and architectures would require minimal changes to the underlying code base. More templates for RTL, power intent, floorplanning and testbench would be required to support new macros and architecture. The current version of the macro flow is written in a python framework used as a wrapper around tcl scripts for synthesis and place & route stage, skill scripts for DRC,

LVS and PEX stages and hspice for post PEX characterization.

7.2.3 Mixed Signal Design Space Exploration

The IVR generation tool enables automated exploration of mixed-signal design space of the IVR and co-optimization of controller, RTL, and physical design thereby extending the scope of design optimization. For example, in a traditional analog design space search, to optimize a design to meet a settling time for fixed inductor, normally F_{SW} and PID gains are controlled. On the other hand, the proposed flow can include digital circuit parameters such as precision of feedback loop for such optimization. To achieve a target settling time for a given inductance, our tool first defines a search space across F_{SW} and inductance. For each F_{SW}/F_{SAMP} in the search space, the tool calculates all possible combinations of precisions for the feedback loop macros which ensure timing closure as shown in Fig. 7.11a. The



Designs	D1	D2	D3	D4	D5	D6	D7	D8
Efficiency (%)	72.54	72.58	73	72.8	72.86	73.1	73.8	73.6
Area (um^2)	29,694.71	29,694.71	29,328.71	29561.16	29561.16	29,328.71	28998.93	28998.93

Figure 7.11: (a) Feedback loop precision characterization with respect to F_{SAMP} (b) Design options for fixed target settling time obtained by optimizing F_{SW}, L and feedback loop precisions for 1.2V-0.8V conversion

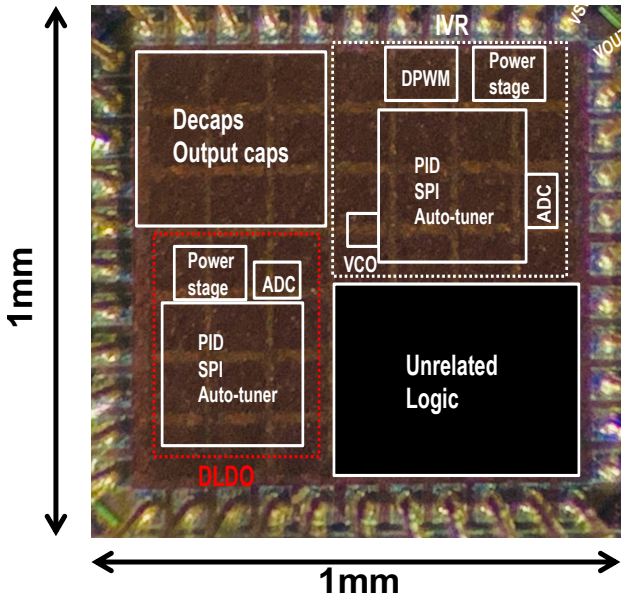


Figure 7.12: 1mm x 1mm Chip Micrograph highlighting the synthesizable IVR and DLDO with essential blocks in 65nm process

tool then converges to multiple design options (Fig. 7.11b) from the search space that meet the target settling time by optimizing inductance, switching frequency and feedback loop precision (obtained from Fig. 7.11a. The final design from the reduced set is selected based on defined constraints and trading off parameters such as efficiency and area.

7.3 Measurement Results

The proposed autogeneration tool from Chapter 6 is used in *generate* mode (Section 6.2.2 and Section 6.2.4) to implement an IVR and DLDO in 65nm CMOS process. The fabricated 1mmx1mm test-chip contains an double sampled IVR with 0.9-1.2V input and 23nF (1.5nF on-die MIM + 1.5nF on-die MOS + 20nF discrete) load capacitor and 62nH (50nH discrete inductor and two bondwires of CLCC44 package estimated at 6nH each) inductance to form output filter. Also in the same 1mmx1mm testchip is a DLDO with input range of 0.6-1.2V with 3.5nF output capacitance. Fig. 7.12 illustrates the chip micrograph & specifications and measurement setup for the designed test chip is shown in Fig. 7.13.

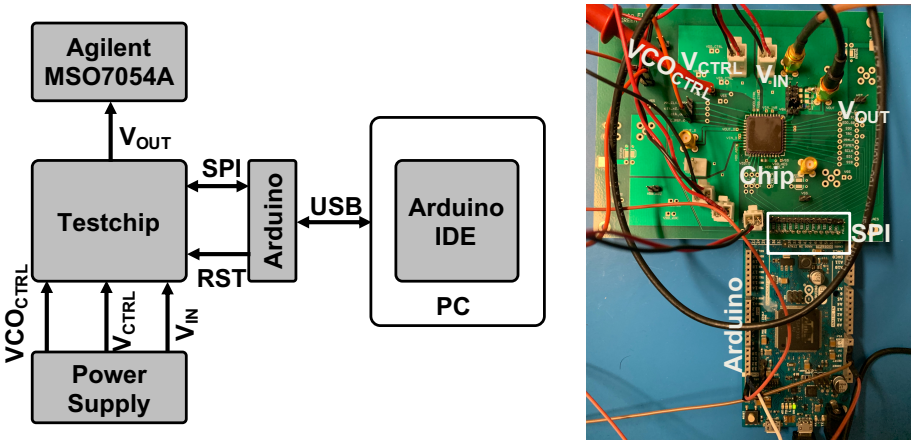


Figure 7.13: Measurement Setup for the test-chip. Arduino micro controller is used to program configurations such as V_{REF}, PID gains, etc to the test-chip from SPI interface and reads out ADC_{OUT}, DPWM control word, etc on a serial monitor terminal

The IVR power stage operates at switching maximum frequency of 120 MHz and is capable of converting 0.9-1.2V input supply to 0.6-1V output range with the ADC resolution of ∼25mV at 5-bit high precision/low-sampling mode. The minimum output is limited by the lower range of the ADC input. Scaling factors are appropriately adjusted to ensure that the scaled outputs are within the ADC range. By default the IVR system operates in HPLS mode where the ADC and the compensators operate at 160-240MHz ($2xF_{SW}$) clock frequency whereas the DPWM converts at the switching frequency of the power stage (80-120MHz). The DLDO always operates at fixed precision.

7.3.1 Macro Characterization

The ADC is characterized at both HPLS and LPHS sampling mode to demonstrate the flexible precision operation. The ADC is characterized by opening the control loop and forcing the V_{OUT} node using an external source and reading out the ADC_{OUT} values via arduino serial monitor interface. Fig. 7.14b illustrates the measured results in both the operating modes. The ADC is tuned to operate in 600-850mV sensing range. Linearity in the sensing range is observed during both the operating modes. Fig. 7.14a illustrates measured characteristics of proposed digital differential VCO. It can be ovserved that the

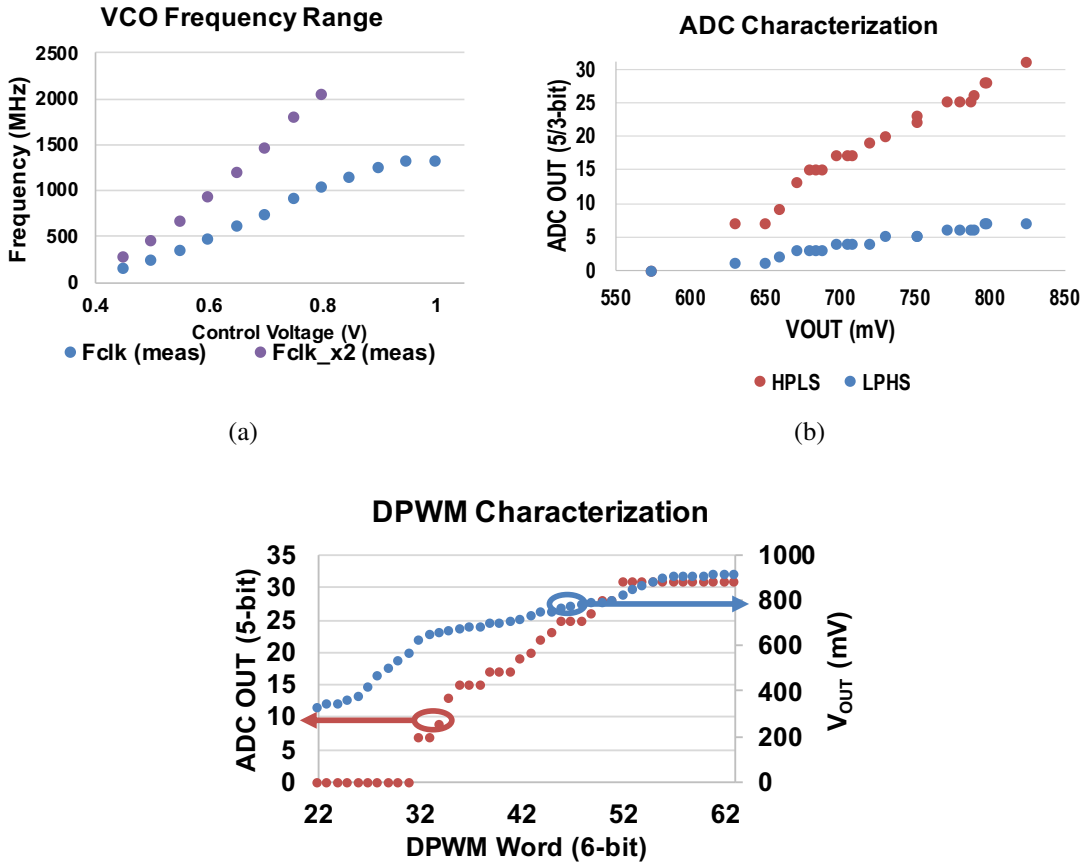


Figure 7.14: Measured results for synthesizeable (a) VCO (b) ADC and (c) DPWM

VCO has near linear frequency range from 125MHz to 1.3GHz for base clock and 250MHz to 2.6GHz for the doubled clock. DPWM is characterized by operating the power stage in the open-loop condition, with varying user definied DPWM input control word in steps of 1 with zero load current and reading out ADC_{OUT} and measuring V_{OUT} levels. The duty cycle variation and V_{OUT} variation with the DPWM control word showed minimal changes when changing the operating modes. This is expected since, the control word is fixed and only the input clock frequency was changing. Fig. 7.14c shows the measured results for DPWM. It can be observed that there is a linear and monotonous increase in V_{OUT} with an increase in the DPWM control word.

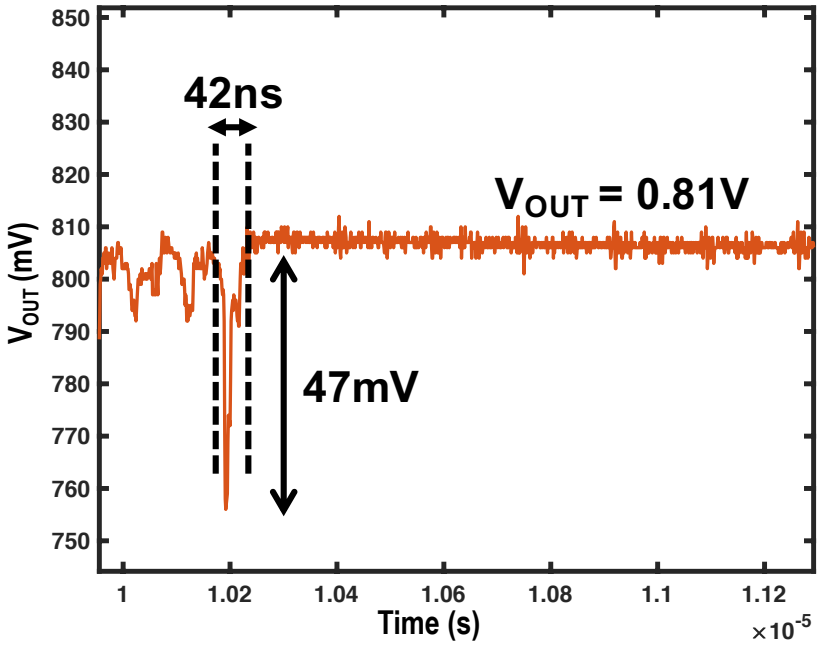


Figure 7.15: Transient response of DLDO operating at V_{IN}=0.88V and V_{OUT}=0.81V under 40mA load jump

7.3.2 DLDO Performance

An on-chip programmable current generator is used to realize fast load steps of varying magnitudes. The 4-bit delay-line ADC runs at 250MHz and control loop utilizes parallel form PID controller for compensation. The programmable load generators with 16 parallel resistances controlled with 16 NMOS switches can generate a maximum load of 70 mA at V_{OUT}=1V. Fig. 7.15 illustrates the measured transient performance of the DLDO operating at V_{IN}=0.88V and V_{OUT}=0.81V under 40mA load jump. The response of load transient demonstrates a recovery/settling time of 42ns for a output droop of 47mV.

7.3.3 IVR Performance

For each operating condition (F_{SW}, output levels, precision mode, etc) the PID coefficients are initially determined by on-chip autotuning. A 30mA/75ps load current step using on-chip load generators and reference step commands corresponding to output level of 650mV

to 780mV are programmed to induce load and reference transients respectively. For the flexible precision operation, when operating at the high-precision/low-sampling (HPLS) mode, the IVR power stage operates at 80MHz and the control loop samples and compensates the output at 160MHz ($2\times F_{SW}$). Whereas during the low-precision/high-sampling (LPHS) mode, the control loop samples and compensates the output and error at even higher rate, i.e. 320MHz ($4\times F_{SW}$). For the variable frequency demonstration, we run the control loop in high-precision mode while changing the switching frequency F_{SW} from 80MHz to 120MHz and corresponding sampling frequency from 160-240MHz.

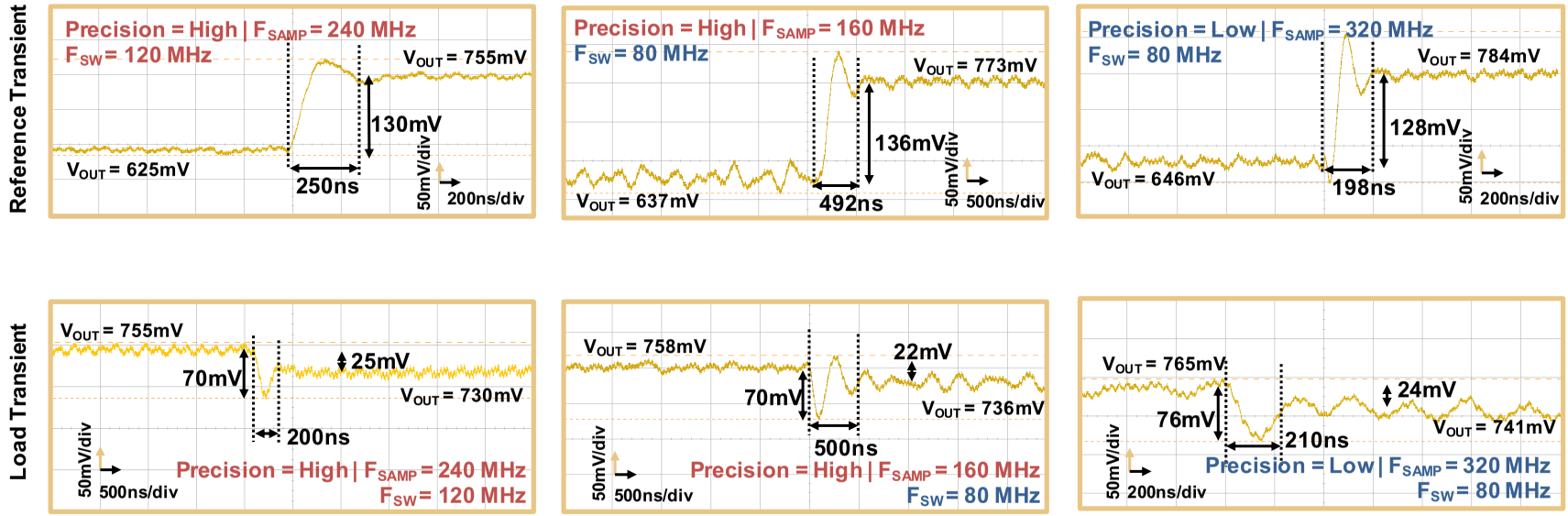
Variable Frequency Operation

Fig. 7.16a and Fig. 7.16b illustrates the measured transient response to reference and load transient events when IVR is operating in high precision HPLS mode. A 70mV droop with a response time of 200ns is observed when operating at F_{SW}=120MHz and F_{SAMP}=240MHz (Fig. 7.16a). The finite (5-bit) ADC resolution results in a 25mV DC drop at V_{OUT} after the load transient. The response to reference transient demonstrates an output slew rate of 0.52V/us at same operating conditions. To demonstrate the variable frequency design, when IVR is operated in same HPLS mode and reference/load steps at F_{SW}=80MHz and F_{SAMP}=160MHz (Fig. 7.16b), the response times increase due to lower F_{SW}.

Flexible Precision Operation

Flexible precision operation is demonstrated by measuring the response time to reference and load steps in different precision modes at fixed F_{SW}=80MHz as illustrated in Fig. 7.16b and Fig. 7.16c. The response times for the same load and reference transients decrease by 60% in the low precision (LPHS, F_{SAMP}=320MHz) mode compared to the high precision (HPLS, F_{SAMP}=160MHz) mode reported in 7.3.3. The higher voltage ripple after the droop (Fig. 7.16c), can be attributed to limit cycling due to ADC using lower precision.

Figure 7.16: Measured results for transient performance of IVR at variable F_{SW} and flexible precision. (a) high precision at F_{SW} = 120MHz; (b) high precision at F_{SW} = 80MHz; (c) low precision at F_{SW} = 80MHz; [**Low:**3-bit ADC / 4-bit PID / F_{SAMP}=4xF_{SW} — **High:**5-bit ADC / 6-bit PID / F_{SAMP}=2xF_{SW}]



(a) (b) (c)

Table 7.1: Comparison with State-of-art DLDOs

Metric	[62] ISSCC'17	[18] TPE'20	[26] TPE'16	[68] JSSC'17	This Work
Technology	40nm	130nm	130nm	28nm	**65nm**
Control Methodology	Digital (Adaptive)	Digital (Adaptive)	Digital (SR)	Digital (SR)	**Digital**
V_{IN} (V)	0.6-1.1	0.5-1.22	0.5-1.2	1.1	**0.6-1.2**
V_{OUT} (V)	0.5-1	0.35-1.17	0.45-1.14	0.9	**0.4-1.13**
Maximum I_L (mA)	210	145	4.6	200	**70**
Load Cap (nF)	20	1.5	1	23.5	**3.5**
Peak Current Efficiency (%)	99.95	97.8	98.3	99.94	**97.4**
Trans. Droop (mV) @ ΔI_L (mA)	36@200	280@40	90@1.4	120@180	**47@40**
Settling Time (ns) @ ΔI_L (mA)	1300@200	55@40	1100@1.4	N/A[a]	**42@40**
Autogenerated	NA	NA	NA	NA	**Yes**

[a] Insufficent Information

7.3.4 Comparasion

The IVR exhibits a 79.3% peak efficiency at 0.78V V_{OUT} and 0.93V V_{IN} at maximum load current of 45mA and the DLDO demonstrates peak current efficieny of 97.4%. Moreover, the analysis shows a good model-hardware correlation between predictions from proposed auto-generation tool and measurements. The measured efficiency and response time to load transients are -8.9% and +18.5% of the modelled (simulated in matlab) design for IVR and +22.4% for response time of DLDO. Table 7.1 and Table 7.2 illustrate that the presented auto-generated DLDO and IVR show competitive performance with prior full/semi-custom designs but with an orders-of-magnitude faster design time.

7.4 Discussion

Discretizing of analog modules is generally preferred because of easier digital implementation and control. However, it results in finite bit-precisions for said modules. In this case, the digitizing of the feedback loop via amplitude quantization of the synthesizable macros leads to non-linear interactions between ADC and DPWM modules. This may result in non-linear effects mainly, limit-cycling oscillations (LCO) [69]. And LCO in turn

Table 7.2: Comparison with State-of-art IVRs

Metric	[14] VLSI'14	[16] ISSCC'17	[6] ISSCC'19	[1] JSSC'17	This Work
Technology	22nm	14nm	28nm	130nm	**65nm**
Control	Digital PWM + Hysteretic	Digital PWM	Digital AOT	Digital PWM + RTA	**Digital PWM Flexible Precision**
Phase	1	2	4	1	**1**
V_{IN} (V)	1.5	21.5	1.2	1.2	**0.9-1.2**
V_{OUT} (V)	1	1.15	0.6-1	0.45-1.05	**0.6-1**
Maximum I_L (mA)	250	330*	1200	70	**45**
F_{SW} (MHz)	500	100	75	125	**80-120**
F_{SAMP} (MHz)	500	100	75	250	**160-240 (High Precision) 320-480 (Low Precision)**
L (nH)	1.5	1.5	15	11.8	**62 (50 + 2xBW)&**
C (nF)	10	50	200	3.2	**23 (20 SMT + 1.5 MIM + 1.5 MOS)**
Peak Efficiency (%)	68@ (V_{OUT}=NA, I_L=90*mA)	84@ (V_{OUT}=1.15, I_L=90*mA)	89@ (V_{OUT}=1, I_L=NA)	71@ (V_{OUT}=0.8, I_L=50*mA)	**79.3@ (V_{OUT}=0.78, I_L=43mA)**
Load Transient Response Time (ns)	NA	NA	80@(V_{OUT}=0.9, ΔI_L=1A)	80@(V_{OUT}=0.9, ΔI_L=60mA)	**200@(V_{OUT}=0.75, ΔI_L=30mA)**
Voltage Ramp	150mV/100ns	500mV/700ns	NA	230mV/80ns	**130mV/250ns**
Synthesizable	Controller	Controller	Controller+	Controller	**Controller + Peripherals**

* Estimated from paper
& Estimated bondwire inductance is 5.8nH
\+ Insufficient Information. Information regarding DAC and power stage are not provided

leads to degradation of static and dynamic regulation performance in digitally controlled IVRs. Usually, LCO can be mitigated by satisfying 1) DPWM having higher resolution than ADC, 2) including non-zero integral gain with an upper limit in control loop and 3) ensuring highest small-signal gain across the ADC for better loop stability. The first and third conditions are usually addressed during the design phase. Overall these conditions generally hold true for traditionally single-sampled systems with a sample-and-hold (SH) ADC. However, multi-sampled systems using traditional SH ADCs can still lead to LCO if the peak-to-peak ripple is mapped to different ADC bins within one switching cycle [1]. To avoid LCO due to this additional condition, instead of traditional SH circuits the proposed synthesizable ADC uses negative edged latches for storage similar to [1]. Additionally, this latched delay line implementation of the proposed ADC also enables capturing any possible changes in the output voltage during the conversion cycle. This is achieved since ADC output depends effectively on averaged output voltage due to different delays of the delay elements during the conversion cycle. This results in same effect as the ADC reported in [1] and other methods such as repetitive-ripple estimation [70] but with no increase in conversion and overall loop delay. Thus, resulting in improved bandwidth and stability for the overall system. Moreover, since auto-tuning cost is accounts for steady-state load conditions, any LCO induced voltage ripple exceeding the ADC bin is captured in the cost and auto-tuning finds the suitable PID coefficients to minimize LCO.

Another major effect of digitizing macros such as VCO is to understand the susceptibility of the synthesized architecture to potential phase noise and jitter. The phase noise and jitter of the VCO can significantly alter the ADC and DPWM characteristics casuing concerns regarding loop stability. In case of the ADC, the change in position of the sampling clock edge can lead to sampling incorrect value of the IVR output. Moreover, for high-bandwith IVRs (switching frequency $>$100MHz), the slew rate further exacerbates the effects of clock jitter. Additionally in case for DPWM, the the jitter and phase noise can lead to generating incorrect duty cycle for a control word. Thus, it is critical to have

a low-jitter architecture for the clock source. For the proposed VCO architecture, the jitter and phase noise can be reduced/mitigated by modifying topology to injection locked ring oscillator as presented in [71] since underlying VCO architecture is similar to the proposed VCO while still being fully synthesizable.

7.5 Summary

This chapter experimentally demonstrates a fully synthesized DLDO and IVR in 65nm testchip implemented using an auto-generation tool flow. The IVR design also includes a flexible precision and variable frequency feedback loop architecture. Synthesizable feedback loop macros accompanied with a macro generation tool flow is also demonstrated to enable the fully synthesizable architecture. The proposed flexible precision feedback loop operating at variable frequency enables trading off switching loss and transient performance. A voltage ramp of 0.52V/µs and peak efficiency of 79.3% are reported for the IVR design. An additionally 60% improvement in transient response is observed when using the flexible precision. And for DLDO, a peak current efficienvy of 97.4% is measured along with fast settling time of 42ns for a 40mA load transient. These results are comparable and competitive to state of the art, full/semi custom designs while enabling orders of magnitude reduction in design time due to the automation.

CHAPTER 8
CONCLUSION AND FUTURE WORK

The reliability and energy efficiency needs in computing systems, ranging from high performance processors to low-power devices are steadily increasing. This thesis details a robust design methodology for reliable and energy efficient self-tuned on-chip voltage regulators, a block primarily used as a solution for maximizing energy/power density and efficiency in modern SoCs. The low PPA (power, performance and area) overhead of the proposed auto-tuning algorithm, easily scalable designs using fully synthesizable architecture and faster design turn around time using the auto-generation tool flow, make the proposed techniques and methodology attractive for implementation. In this chapter, we walk through a summary of the main contributions of this thesis in Section 8.1. We conclude by examining future research directions in Section 8.2.

8.1 Dissertation Summary

This thesis starts with identifying various challenges regarding self-tuning of inductive IVR and mainly focuses on co-tuning of the IVR with the digital core since they are fabricated on the same die and will incur similar variations. **Chapter 3** demonstrates through a simulation framework that tuning inductive IVR in isolation does improve the transient performance of the regulator, but the performance of the digital core may not always be optimal. Thus, a tuning metric is needed to be defined to account for overall system performance. **Chapter 3** concludes that performance based tuning of an IVR can be used to improve both the transient performacne of the IVR and the performance of digital core. A maximum of 12.18% (33.98MHz) of operating frequency improvement was observed in simulations using the proposed performance based co-tuning.

Simulation based results might underestimate or overestimate the improvement in per-

formance of the digital core, particularly because of inherent mismatches between models, variations and actual hardware implementation. Therefore, it is extremely crucial to validate the improvement in performance improvement of the digital core through a hardware prototype. **Chapter 4** identifies the design issues of translating the proposed performance based tuning algorithm into lightweight hardware and details measurement results of an all-digital architecture of an inductive IVR driving an AES core using package bondwires as inductances, implemented in 130nm CMOS process and tuning against process & passive variations using on-chip delay sensors. The designed system along with the proposed performance based tuning showcased 5.2% (4.16MHz) improvement in the maximum operating frequency of the AES engine.

Chapter 5 characterizes the effects of NBTI induced aging degradations in IVR and DLDOs while focusing on two main regions: Power stage and feedback loop controller. The IVR system shows minimal effects of power stage aging on the transient behaviour, but does showcase a slight drop in power efficiency. Whereas DLDO measurement and simulation results show significant (upto 71.4% for 65nm test-chip) degradation in transient performance due to power stage aging. Additionally, the feedback loop controller aging simulations demonstrate upto 13.9% and 13.1% degradation in transient performance for DLDO and IVR respectively. Moreover, the measurements show that auto-tuning of the DLDO can improve the transient response by upto 30%. This further reinforces need for auto-tuning of on-chip voltage regulator designs.

Chapter 6 introduces an auto-generation tool flow for high bandwidth on-chip voltage regulators to reduce the design time and improve scalability. The auto-generation tool is capable of designing IVR and DLDO with fixed specifications or design an optimal IVR or DLDO based on some specifications and constraints by optimizing control loop and power stage design. The designed/auto-generated VRs showcase comparable specification to state of art custom/semi-custom designs while reducing the design time in orders of magnitude. The modular nature of the tool allows for even faster runtime by using a better and more

advanced optimization function.

Chapter 7 explores a fully synthesizable architecture of an IVR and DLDO to further simplify the integration with the auto-generation tool and scalability to advanced process nodes. New synthesizable architectures for feedback loop macros are also presented along with a macro generation flow to seamlessly integrate with auto-generation tool flow. Additionally, a flexible precision feedback loop has been demonstrated in IVR to improve transient performance by trading off bit-precision and accuracy with higher sampling rate. Measurement results of the designed prototype chip demonstrates a peak efficiency of 79.3%, 0.52V/µs voltage ramp and upto 60% improvement in transient response using the flexible precision architecture for IVR and a 97.4% current efficiency for DLDO along with fast response time of 42ns for a 40mA load transient.

8.2 Future Directions

Most of the contributions of this thesis can be implemented immediately to practical applications. Additionally, the findings from this thesis can be extended in several potential directions for future research.

Chapter 3 offers insight into impact of co-tuning the IVR along with the digital core. However, a major challenge in tuning of on-chip regulators in more practical large scale applications is tuning for systems with distributed power delivery architecture. The challenge in tuning of a distributed power delivery system with multiple IVRs and DLDOs is to characterize the effect of cross coupling between multiple VRMs. Consider a distributed power delivery architecture with a global FIVR and multiple local DLDOs for point-of-load regulation. If the load serviced by one DLDO makes a transition, it injects noise into its input power line (i.e. output of global IVR), which appears as power supply noise for other DLDOs. Hence, the load generation mechanism and the tuning engine needs to consider the cross-coupled noise to characterize stability and output voltage response. The tuning for the distributed DLDOs is more complicated as we need to consider the cross-coupled

noise and develop a load-step scheduling approach for worst-case cross-coupling noise. The worst-case cross-coupled noise for a DLDO under test may occur when the load step at that DLDO is applied after a certain delay from the load step of all other DLDOs. The finite delay allows the noise generated at the power supply node of the other DLDOs to propagate to the power supply node of the DLDO under test.

The auto-generation tool flow discussed in Chapter 6 also opens up a wide range of possible extensions for future research. Currently, the proposed tool supports just a single phase VR architecture with a linear control loop. Since most of the modern state of the art regulators implement multi-phase architecture for better efficiency and non-linear control loop elements such as resistive transient assists [1] for improved transient performance, supporting multiple architectures of various on-chip regulators would be one of the most useful features. Moreover, making the tool capable of determining which converter and/or architecture would be most appropriate based on a high level input specifications and constraints, leading to creation a collective database of models, control techniques and template layouts would significantly enhance the usability of the tool.

Another aspect to focus from the tool point of view can be integration with a digital SoC. The auto-generation flow must integrate physical design of the IVR within the SoC at different levels of granularity. For example, the simplest option, as demonstrated in Fig. 6.10, is to integrate a single IVR as a hard macro within the SoC physical design flow and connect the output of the IVR to the SoC power grid. However, we can place the IVR close to the higher power blocks within the SoC to reduce power supply noise. Additionally, a more efficient, but complex, approach is to optimally distribute/place the IVR power stages (and output capacitors) within a digital block (or SoC) to reduce supply noise. In this case, the entire IVR is not considered as a hard macro, rather, the power stages are distributed while feedback path is considered as a macro.

REFERENCES

[1] M. Kar, A. Singh, A. Rajan, V. De, and S. Mukhopadhyay, "An All-Digital Fully Integrated Inductive Buck Regulator With A 250-MHz Multi-Sampled Compensator and a Lightweight Auto-Tuner in 130-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 7, pp. 1825–1835, Jul. 2017.

[2] J. A. A. Qahouq and V. Arikatla, "Online Closed-Loop Autotuning Digital Controller for Switching Power Converters," *IEEE Transactions on Industrial Electronics*, vol. 60, no. 5, pp. 1747–1758, May 2013.

[3] M. Shirazi, R. Zane, and D. Maksimovic, "An Autotuning Digital Controller for DC-DC Power Converters Based on Online Frequency-Response Measurement," *IEEE Transactions on Power Electronics*, vol. 24, no. 11, pp. 2578–2588, Nov. 2009.

[4] H. Luo, H. Li, L. Yeh, and C. J. Liu, "Automated synthesis design flow of power converter circuits aimed at soc applications," in *2011 International Symposium on Integrated Circuits*, Dec. 2011, pp. 281–284.

[5] H. Hsu, W. Chen, L. Yeh, and C. Jimmy Liu, "Spec-to-layout automation flow for buck converters with current-mode control in soc applications," in *2018 15th International Conference on Synthesis, Modeling, Analysis and Simulation Methods and Applications to Circuit Design (SMACD)*, Jul. 2018, pp. 169–172.

[6] M. Choi, C. Kye, J. Oh, M. Choo, and D. Jeong, "27.7 a synthesizable digital aot 4-phase buck voltage regulator for digital systems with 0.0054mm2 controller and 80ns recovery time," in *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, Feb. 2019, pp. 432–434.

[7] E. A. Burton, G. Schrom, F. Paillet, J. Douglas, W. J. Lambert, K. Radhakrishnan, and M. J. Hill, "FIVR − Fully integrated voltage regulators on 4th generation Intel® Core™ SoCs," in *2014 IEEE Applied Power Electronics Conference and Exposition - APEC 2014*, Mar. 2014, pp. 432–439.

[8] E. J. Fluhr, J. Friedrich, D. Dreps, V. Zyuban, G. Still, C. Gonzalez, A. Hall, D. Hogenmiller, F. Malgioglio, R. Nett, J. Paredes, J. Pille, D. Plass, R. Puri, P. Restle, D. Shan, K. Stawiasz, Z. T. Deniz, D. Wendel, and M. Ziegler, in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), title=5.1 POWER8™: A 12-core server-class processor in 22nm SOI with 7.6Tb/s off-chip bandwidth*, 2014, pp. 96–97.

[9] N. Sturcken, M. Petracca, S. Warren, P. Mantovani, L. P. Carloni, A. V. Peterchev, and K. L. Shepard, "A Switched-Inductor Integrated Voltage Regulator With Nonlinear Feedback and Network-on-Chip Load in 45 nm SOI," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 8, pp. 1935–1945, Aug. 2012.

[10] S. S. Kudva and R. Harjani, "Fully-Integrated On-Chip DC-DC Converter With a 450X Output Range," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 8, pp. 1940–1951, Aug. 2011.

[11] W. Kim, D. Brooks, and G. Y. Wei, "A Fully-Integrated 3-Level DC-DC Converter for Nanosecond-Scale DVFS," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 1, pp. 206–219, Jan. 2012.

[12] S. T. Kim, Y. C. Shih, K. Mazumdar, R. Jain, J. F. Ryan, C. Tokunaga, C. Augustine, J. P. Kulkarni, K. Ravichandran, J. W. Tschanz, M. M. Khellah, and V. De, "Enabling Wide Autonomous DVFS in a 22 nm Graphics Execution Core Using a Digitally Controlled Fully Integrated Voltage Regulator," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 18–30, Jan. 2016.

[13] M. Lee, Y. Choi, and J. Kim, "A 500-mhz, 0.76-w/mm power density and 76.2% power efficiency, fully integrated digital buck converter in 65-nm cmos," *IEEE Transactions on Industry Applications*, vol. 52, no. 4, pp. 3315–3323, Jul. 2016.

[14] H. K. Krishnamurthy, V. A. Vaidya, P. Kumar, G. E. Matthew, S. Weng, B. Thiruvengadam, W. Proefrock, K. Ravichandran, and V. De, "A 500 mhz, 68% efficient, fully on-die digitally controlled buck voltage regulator on 22nm tri-gate cmos," in *2014 Symposium on VLSI Circuits Digest of Technical Papers*, Jun. 2014, pp. 1–2.

[15] C. Huang and P. K. T. Mok, "An 84.7% Efficiency 100-MHz Package Bondwire-Based Fully Integrated Buck Converter With Precise DCM Operation and Enhanced Light-Load Efficiency," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2595–2607, Nov. 2013.

[16] H. K. Krishnamurthy, V. Vaidya, S. Weng, K. Ravichandran, P. Kumar, S. Kim, R. Jain, G. Matthew, J. Tschanz, and V. De, "20.1 A digitally controlled fully integrated voltage regulator with on-die solenoid inductor with planar magnetic core in 14nm tri-gate CMOS," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2017, pp. 336–337.

[17] S. B. Nasir, S. Gangopadhyay, and A. Raychowdhury, "5.6 A 0.13µm fully digital low-dropout regulator with adaptive control and reduced dynamic stability for ultra-wide dynamic range," in *2015 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2015, pp. 1–3.

[18] A. Singh, M. Kar, V. C. K. Chekuri, S. Mathew, A. Rajan, V. De, and S. Mukhopadhyay, "A digital low-dropout regulator with auto-tuned pid compensator and dynamic gain control for improved transient performance under process variations and aging," *IEEE Transactions on Power Electronics*, pp. 1–1, 2019.

[19] N. Sturcken, E. O'Sullivan, N. Wang, P. Herget, B. Webb, L. Romankiw, M. Petracca, R. Davies, R. Fontana, G. Decad, I. Kymissis, A. Peterchev, L. Carloni, W. Gallagher, and K. Shepard, "A 2.5D integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer delivering 10.8A/mm2," in *2012 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2012, pp. 400–402.

[20] S. Arora, D. K. Su, and B. A. Wooley, "A compact 120-mhz 1.8v/1.2v dual-output dc-dc converter with digital control," in *Proceedings of the IEEE 2013 Custom Integrated Circuits Conference*, Sep. 2013, pp. 1–4.

[21] V. C. K. Chekuri, M. Kar, A. Singh, A. K. Davis, M. L. F. Bellaredj, M. Swaminathan, and S. Mukhopadhyay, "An Inductive Voltage Regulator with Overdrive Tracking across Input Voltage in Cascoded Power Stage," *IEEE Transactions on Circuits and Systems II: Express Briefs*, pp. 1–1, 2020.

[22] S. Bandyopadhyay, Y. K. Ramadass, and A. P. Chandrakasan, "20 μA to 100 mA DC−DC Converter With 2.8-4.2 V Battery Supply for Portable Applications in 45 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 12, pp. 2807–2820, Dec. 2011.

[23] S. J. Kim, R. K. Nandwana, Q. Khan, R. C. N. Pilawa-Podgurski, and P. K. Hanumolu, "A 4-Phase 30−70 MHz Switching Frequency Buck Converter Using a Time-Based Compensator," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 12, pp. 2814–2824, Dec. 2015.

[24] P. Hammarlund, A. J. Martinez, A. A. Bajwa, D. L. Hill, E. Hallnor, H. Jiang, M. Dixon, M. Derr, M. Hunsaker, R. Kumar, R. B. Osborne, R. Rajwar, R. Singhal, R. D'Sa, R. Chappell, S. Kaushik, S. Chennupaty, S. Jourdan, S. Gunther, T. Piazza, and T. Burton, "Haswell: The Fourth-Generation Intel Core Processor," *IEEE Micro*, vol. 34, no. 2, pp. 6–20, Mar. 2014.

[25] B. Keller, M. Cochet, B. Zimmer, Y. Lee, M. Blagojevic, J. Kwak, A. Puggelli, S. Bailey, P. F. Chiu, P. Dabbelt, C. Schmidt, E. Alon, K. Asanović, and B. Nikolić, "Sub-microsecond adaptive voltage scaling in a 28nm FD-SOI processor SoC," in *ESSCIRC Conference 2016: 42nd IEEE European Solid-State Circuits Conference*, Sep. 2016, pp. 269–272.

[26] S. B. Nasir, S. Gangopadhyay, and A. Raychowdhury, "All-digital low-dropout regulator with adaptive control and reduced dynamic stability for digital load circuits," *IEEE Transactions on Power Electronics*, vol. 31, no. 12, pp. 8293–8302, 2016.

[27] X. Ma, Y. Lu, R. P. Martins, and Q. Li, "A 0.4v 430na quiescent current nmos digital ldo with nand-based analog-assisted loop in 28nm cmos," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 306–308.

[28] M. Huang, Y. Lu, U. Seng-Pan, and R. P. Martins, "20.4 an output-capacitor-free analog-assisted digital low-dropout regulator with tri-loop control," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 342–343.

[29] T. Singh, S. Rangarajan, D. John, C. Henrion, S. Southard, H. McIntyre, A. Novak, S. Kosonocky, R. Jotwani, A. Schaefer, E. Chang, J. Bell, and M. Co, "3.2 zen: A next-generation high-performance x86 core," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 52–53.

[30] M. Cho, S. T. Kim, C. Tokunaga, C. Augustine, J. P. Kulkarni, K. Ravichandran, J. W. Tschanz, M. M. Khellah, and V. De, "Postsilicon voltage guard-band reduction in a 22 nm graphics execution core using adaptive voltage scaling and dynamic power gating," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 50–63, Jan. 2017.

[31] K. Bowman, J. Tschanz, C. Wilkerson, S.-L. Lu, T. Karnik, V. De, and S. Borkar, "Circuit techniques for dynamic variation tolerance," in *2009 46th ACM/IEEE Design Automation Conference (DAC)*, Jul. 2009, pp. 4–7.

[32] J. Tschanz, N. S. Kim, S. Dighe, J. Howard, G. Ruhl, S. Vangal, S. Narendra, Y. Hoskote, H. Wilson, C. Lam, M. Shuman, C. Tokunaga, D. Somasekhar, S. Tang, D. Finan, T. Karnik, N. Borkar, N. Kurd, and V. De, "Adaptive Frequency and Biasing Techniques for Tolerance to Dynamic Temperature-Voltage Variations and Aging," in *2007 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, Feb. 2007, pp. 292–604.

[33] K. A. Bowman, C. Tokunaga, T. Karnik, V. K. De, and J. W. Tschanz, "A 22 nm All-Digital Dynamically Adaptive Clock Distribution for Supply Voltage Droop Tolerance," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 4, pp. 907–916, Apr. 2013.

[34] M. Kar, S. Carlo, H. Krishnamurthy, and S. Mukhopadhyay, "Impact of process variation in inductive Integrated Voltage Regulator on delay and power of digital circuits," in *2014 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED)*, Aug. 2014, pp. 227–232.

[35] N. Sturcken, R. Davies, H. Wu, M. Lekas, K. Shepard, K. W. Cheng, C. C. Chen, Y. S. Su, C. Y. Tsai, K. D. Wu, J. Y. Wu, Y. C. Wang, K. C. Liu, C. C. Hsu, C. L. Chang, W. C. Hua, and A. Kalnitsky, "Magnetic thin-film inductors for monolithic integration with CMOS," in *2015 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2015, pp. 11.4.1–11.4.4.

[36] S. Das, C. Tokunaga, S. Pant, W. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. T. Blaauw, "Razorii: In situ error detection and correction for pvt and ser tolerance," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[37] D. Ernst, Nam Sung Kim, S. Das, S. Pant, R. Rao, Toan Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, and T. Mudge, "Razor: A low-power pipeline based on circuit-level timing speculation," in *Proceedings. 36th Annual IEEE/ACM International Symposium on Microarchitecture, 2003. MICRO-36.*, Dec. 2003, pp. 7–18.

[38] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D. M. Harris, D. Blaauw, and D. Sylvester, "Bubble razor: Eliminating timing margins in an arm cortex-m3 processor in 45 nm cmos using architecturally independent error detection and correction," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 66–81, Jan. 2013.

[39] A. Costabeber, P. Mattavelli, S. Saggini, and A. Bianco, "Digital autotuning of dc-dc converters based on model reference impulse response," in *2010 Twenty-Fifth Annual IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2010, pp. 1287–1294.

[40] ——, "Digital autotuning of dc–dc converters based on a model reference impulse response," *IEEE Transactions on Power Electronics*, vol. 26, no. 10, pp. 2915–2924, 2011.

[41] W. Stefanutti, P. Mattavelli, S. Saggini, and M. Ghioni, "Autotuning of digitally controlled buck converters based on relay feedback," in *2005 IEEE 36th Power Electronics Specialists Conference*, 2005, pp. 2140–2145.

[42] S. Saggini, A. Costabeber, and P. Mattavelli, "A simple digital autotuning for analog controller in smps," *IEEE Transactions on Power Electronics*, vol. 25, no. 8, pp. 2170–2178, 2010.

[43] W. Deng, D. Yang, T. Ueno, T. Siriburanon, S. Kondo, K. Okada, and A. Matsuzawa, "A fully synthesizable all-digital pll with interpolative phase coupled oscillator, current-output dac, and fine-resolution digital varactor using gated edge injection technique," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 1, pp. 68–80, Jan. 2015.

[44] B. Xu, S. Li, N. Sun, and D. Z. Pan, "A scaling compatible, synthesis friendly vco-based delta-sigma adc design and synthesis methodology," in *2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, Jun. 2017, pp. 1–6.

[45] S. Henzler, *Time-to-Digital Converters*, 1st. Springer Publishing Company, Incorporated, 2010, ISBN: 9048186277, 9789048186273.

[46] D. Fick, N. Liu, Z. Foo, M. Fojtik, J. s. Seo, D. Sylvester, and D. Blaauw, "In situ delay-slack monitor for high-performance processors using an all-digital self-calibrating 5ps resolution time-to-digital converter," in *2010 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2010, pp. 188–189.

[47] *MOSIS Packaging Service*, Available: `https://www.mosis.com/pages/products/assembly/index` (2020-07-18).

[48] M. Alam and S. Mahapatra, "A comprehensive model of pmos nbti degradation," *Microelectronics Reliability*, vol. 45, no. 1, pp. 71 –81, 2005.

[49] R. Vattikonda, W. Wang, and Y. Cao, "Modeling and minimization of pmos nbti effect for robust nanometer design," in *Proceedings of the 43rd Annual Design Automation Conference*, ser. DAC '06, Association for Computing Machinery, 2006, 1047–1052, ISBN: 1595933816.

[50] J. Keane, T. Kim, and C. H. Kim, "An on-chip nbti sensor for measuring pmos threshold voltage degradation," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 18, no. 6, pp. 947–956, Jun. 2010.

[51] J. Wu, A. Boyer, J. Li, S. B. Dhia, and R. Shen, "Characterization of changes in ldo susceptibility after electrical stress," *IEEE Transactions on Electromagnetic Compatibility*, vol. 55, no. 5, pp. 883–890, 2013.

[52] L. Wang, S. K. Khatamifard, U. R. Karpuzcu, and S. Köse, "Mitigation of nbti induced performance degradation in on-chip digital ldos," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, 2018, pp. 803–808.

[53] V. C. K. Chekuri, A. Singh, N. Dasari, and S. Mukhopadhyay, "On the effect of nbti induced aging of power stage on the transient performance of on-chip voltage regulators," in *2019 IEEE International Reliability Physics Symposium (IRPS)*, Mar. 2019, pp. 1–5.

[54] V. C. K. Chekuri, N. M. Rahman, E. Lee, A. Signh, and S. Mukhopadhyay, "A Fully Synthesized Integrated Buck Regulator with Auto-generated GDS-II in 65nm CMOS Process," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*, 2020, pp. 1–4.

[55] V. C. K. Chekuri, M. Kar, A. Singh, and S. Mukhopadhyay, "Autotuning of integrated inductive voltage regulator using on-chip delay sensor to tolerate process and passive variations," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 8, pp. 1768–1778, Aug. 2019.

[56] L. Corradini, D. Maksimovic, P. Mattavelli, and R. Zane, *Digital Control of High-Frequency Switched-Mode Power Converters*. Wiley-IEEE Press, 2015.

[57] S. Mueller, K. Z. Ahmed, A. Singh, A. K. Davis, S. Mukhopadyay, M. Swaminathan, Y. Mano, Y. Wang, J. Wong, S. Bharathi, H. F. Moghadam, and D. Draper, "Design of high efficiency integrated voltage regulators with embedded magnetic core inductors," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, May 2016, pp. 566–573.

[58] M. Lee, A. Singh, H. M. Torun, J. Kim, S. Lim, M. Swaminathan, and S. Mukhopadhyay, "Automated generation of all-digital i/0 library cells for system-in-package integration of multiple dies," in *2018 IEEE 27th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)*, Oct. 2018, pp. 65–67.

[59] V. Unnikrishnan and M. Vesterbacka, "Time-mode analog-to-digital conversion using standard cells," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 12, pp. 3348–3357, Dec. 2014.

[60] R. Franch, P. Restle, N. James, W. Huott, J. Friedrich, R. Dixon, S. Weitzel, K. Van Goor, and G. Salem, "On-chip timing uncertainty measurements on ibm microprocessors," in *2008 IEEE International Test Conference*, Oct. 2008, pp. 1–7.

[61] V. Yousefzadeh, T. Takayama, and D. Maksimovi, "hybrid dpwm with digital delay-locked loop," in *2006 IEEE Workshops on Computers in Power Electronics*.

[62] W. Tsou, W. Yang, J. Lin, H. Chen, K. Chen, C. Wey, Y. Lin, S. Lin, and T. Tsai, "20.2 digital low-dropout regulator with anti pvt-variation technique for dynamic voltage scaling and adaptive voltage scaling multicore processor," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*.

[63] *IVR-Gen*, Available: `https://github.com/GT-CHIPS/IVR-Gen.git` (2020-07-18).

[64] V. C. K. Chekuri, N. Dasari, A. Singh, and S. Mukhopadhyay, "Automatic gdsii generator for on-chip voltage regulator for easy integration in digital socs," in *2019 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Jul. 2019, pp. 1–6.

[65] V. C. K. Chekuri, M. Kar, A. Singh, and S. Mukhopadhyay, "Performance based tuning of an inductive integrated voltage regulator driving a digital core against process and passive variations," in *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*, Mar. 2018, pp. 367–372.

[66] V. C. K. Chekuri, A. Singh, N. M. Rahman, E. Lee, and S. Mukhopadhyay, "Aging challenges in on-chip voltage regulator design," in *2020 IEEE International Reliability Physics Symposium (IRPS)*, 2020, pp. 1–8.

[67] A. Drake, R. Senger, H. Deogun, G. Carpenter, S. Ghiasi, T. Nguyen, N. James, M. Floyd, and V. Pokala, "A Distributed Critical-Path Timing Monitor for a 65nm High-Performance Microprocessor," in *2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers*, 2007.

[68] Y. Lee, W. Qu, S. Singh, D. Kim, K. Kim, S. Kim, J. Park, and G. Cho, "A 200-ma digital low drop-out regulator with coarse-fine dual loop in mobile application processor," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 64–76, 2017.

[69] L. Corradini, D. Maksimovic, P. Mattavelli, and R. Zane, "Amplitude quantization," in *Digital Control of High-Frequency Switched-Mode Power Converters*. Wiley-IEEE Press, 2015, ch. 5, pp. 167–190.

[70] L. Corradini, P. Mattavelli, E. Tedeschi, and D. Trevisan, "High-bandwidth multisampled digitally controlled dc–dc converters using ripple compensation," *IEEE Transactions on Industrial Electronics*, vol. 55, no. 4, pp. 1501–1508, 2008.

[71] N. T. Abou-El-Kheir, R. D. Mason, M. Li, and M. C. E. Yagoub, "A 65 nm compact high performance fully synthesizable clock multiplier based on an injection locked ring oscillator," in *2018 IEEE International Symposium for Circuits and Systems (ISCAS)*, 2018, pp. 1–5.